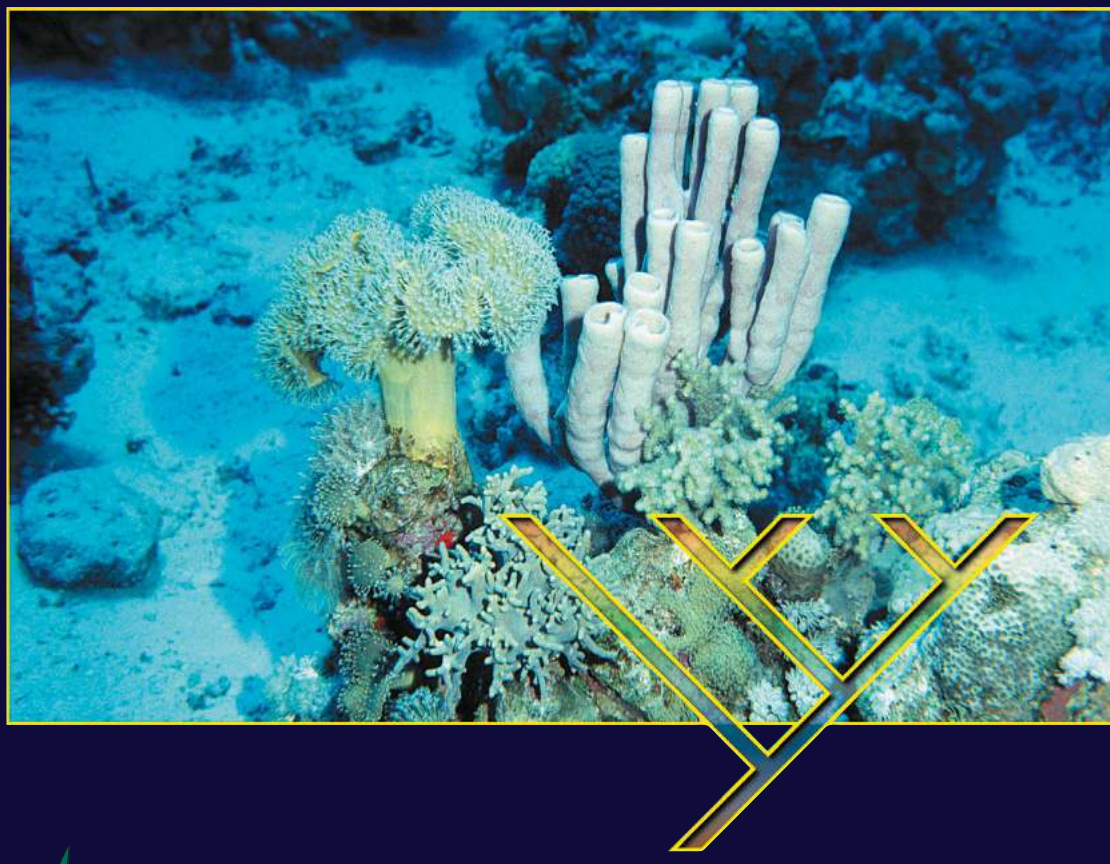
Foundations of Phylogenetic Systematics

Johann-Wolfgang Wägele



Verlag Dr. Friedrich Pfeil • München

Johann-Wolfgang Wägele

Foundations of Phylogenetic Systematics

Bibliographic Information published by Die Deutschen Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at
http://dnb.ddb.de.

Original Title:
Grundlagen der Phylogenetischen Systematik,
first published by Verlag Dr. Friedrich Pfeil, München, 2000
2nd, revised edition 2001.

Druckvorstufe: Verlag Dr. Friedrich Pfeil, München
CTP-Druck: grafik + druck GmbH Peter Pöllinger, München
Buchbinder: Thomas, Augsburg

ISBN 3-89937-056-2

Printed in Germany

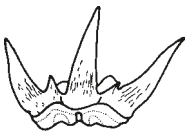
Verlag Dr. Friedrich Pfeil, Wolfratshauser Straße 27, D-81379 München, Germany
Tel.: +49-(0)89-74 28 270 • Fax: +49-(0)89-72 42 772 • E-Mail: info@pfeil-verlag.de
www.pfeil-verlag.de

Foundations of

Phylogenetic Systematics

Johann-Wolfgang Wägele

Translated from the German second edition
by C. Stefen, J.-W. Wägele, and revised by B. Sinclair



Verlag Dr. Friedrich Pfeil
München, January 2005
ISBN 3-89937-056-2

Contents

Introduction

The central tasks of systematics are the inventory, systematization and description of the diversity of organisms which evolved during the earth's history (the term "systematization" is defined in ch. 1.3.4). With the assignment of an organism to a taxon, statements concerning its anatomy and its role and trophic position in ecosystems are organized. This systematization facilitates the management of the genetic information present in nature, which can be considered to be a collection of blueprints of biodiversity, the inestimably valuable and currently light-heartedly wasted inheritance of humankind. The systematist creates a reference system that not only allows direct retrieval of biological information, but also predictions about the properties of organisms. Ecological and evolutionary research rely on the results of systematics.

Phylogenetic systematics (= **phylogenetics**) deals with the detection and substantiation of relationships, enabling an intersubjectively* testable placement of taxa in a phylogenetic tree. The aim is to depict the phylogenesis (= phylogeny), the historical series of divergence events that biologists usually call the evolution of species (but what is a species? see ch. 2.3), in tree graphs. In this book the term "**evolution**" is used free of anthropocentric prejudices in the sense of "change in the course of time", the visible result of processes that modify the anatomy and lifestyle of organisms through time leaving traces in the organisms' information-coding molecules. These processes not only modify adaptive characters and genes, but also other characters including some DNA sequences which seemingly have no function.

In the last years, systematics gained increasing appreciation as empirical science, because it became clearer that objective probability decisions are made, which can partly be described with mathematical functions. In the past the instability of taxon names and of the classification of organisms, often the result of subjective decisions that are scientifically hardly justifiable, created an unfavourable impression of the efficiency and importance of systematics. In recent times, the results obtained with "modern methods" increased the confusion, because many publications were based on insufficient data and unsound methods of data analysis. New hypotheses were accepted by many without the necessary scepticism. Disagreements between systematists helped to nourish doubts about the objectivity of the methods and the testability of hypotheses used in systematics, properties a hard science should have. The reference to a well-founded theory of systematics was often lacking.

It is the goal of the following chapters to present the theoretical basis of objective data analysis. The same basic laws of systematics are valid for both comparative morphology and the analysis of DNA sequences. It is the intention of this book to present a comparison and synthesis of the methodological procedure of Hennigian phylogenetic systematics with new numerical methods, to depict what these methods have in common and where they differ, and to search for a common theoretical basis. Basic theories of phylogenetic systematics were originally developed by the talented entomologist and theorist W. Hennig (1913–1976). Today the repertoire of this science includes cladistics and many other approaches.

A profound knowledge of the theory is an essential prerequisite for scientific work. Additionally, experience is essential. This can only be gained through analyses of real examples and not through reading alone. Only working with real organisms and with data obtained from analyses of nature, will one understand the peculiarities of methods and also of individual groups of organisms.

The natural historical events which produced the patterns of correspondences in construction and lifestyle that we observe when comparing living

* *intersubjectively* means that different persons would make independently the same observations or they would arrive at the same conclusions (see also ch. 1.4.2).

organisms, especially the multitude of evolutionary processes that led to the genetic divergence of populations, do not have to be explored in detail to reconstruct phylogeny. In most cases these processes are not known in detail anyway. Especially interesting are those processes leading to *irreversible* splitting of populations. However, it is essential for a systematist to know the principle mechanisms that change the structure of organisms over time, in order to be able to decide which methods can be used to analyse the data at hand; if it were known that a character evolves at a constant rate, the time elapsed since a divergence event could be determined using the observable character states. For this reason, chapters on the evolution of characters are included. The processes leading to the substitution of an old character state by a younger one take place in populations. The importance of population genetics is stressed in this context, but this field of biology cannot be treated in this book.

By now there exist a large and not easily surveyed number of methods proposed to identify evolutionary novelties that evolved in organisms and to reconstruct their phylogeny. It is not within the scope of this book to discuss them all. Furthermore, it is not always rewarding to invest time in the testing and application of new methods if their development is not concluded. It is, however, essential to show the epistemological basis of systematics, which is valid for all methods and which should be understood by each systematist. The methods of phylogenetic systematics are still evolving, however, the general and strict logical foundations that can be inferred theoretically are invariable.

Phylogeny (= cladogenesis, phylogenesis)**:** the natural process of repeated splitting of populations through irreversible genetic divergence.

Biological systematics: science of the systematization of organisms and of the description of their genetic and phenotypical diversity (= biodiversity).

Phylogenetic systematics: detection and substantiation of phylogenetic relationships of groups of organisms, and integration of proper names of groups of organisms into a mental system that reflects their phylogeny (see term "systematization", ch. 1.3.4).

Phylogenetics: science of the reconstruction of phylogeny (subdiscipline of phylogenetic systematics).

Cladistics: construction of dendrograms from character/taxa datasets using the maximum parsimony method (one of several available methods; see ch. 1.4.5).

Acknowledgements: In the course of seminars on phylogenetic systematics and during everyday conversations several ideas arose which complete this text. In particular I thank my wife Heike, the members of my lab, friends and colleagues:

Oliver Coleman, Hermann Dreyer, Ulrike Englisch, Martin Fanenbruck, Christoph Held, Wulf Kobusch, Andreas Leistikow, Friederike Rödding, Christian Schmidt, Günter Stanjek, Evi Wollscheid.

For technical support in the production of the figures, I would like to thank Steffen Koehler, Andrea Kogelheide, Ingo Manstedt und Ilse Weßel. The author is also grateful to C. Stefen and Dr. B. Sinclair for help with the translation and to the publisher Dr. Pfeil for his patience with the preparation of the manuscript.

Note: the author would be grateful for any comment that can improve the text.

Bochum, October 2004
Johann-Wolfgang Wägele

1. Metaphysical foundations of science

Natural sciences aim to perceive and describe existing facts. Laws deduced from observations allow us to make predictions which are the basis of recommendations for future action. Scientific "observations" of our surroundings form the picture we have of our environment. Epistemology (the theory of science) explains why we can have some certainty that this picture equals reality and how scientific knowledge is gained.

Biologists dealing with systematics will sooner or later encounter the question of whether hypotheses of relationships, named species, or reconstructed evolutionary events represent the result of real processes, or if they are material objects of reality, or whether they are a product of our fallacies. For an outsider, the incompatibility of viewpoints and hypotheses of different systematists (Fig. 54, 182) may indicate subjectivity and the lack of a logically founded scientific basis. Such a discipline would not be taken seriously as a hard science. In fact, there are gaps in our methodology which give rise to criticism. For example, the decision whether a similarity is a homology or an analogy is often made "due to feelings" or with reference to personal experience. Elegant numerical methods have been developed and employed for several publications without checking the assumptions these methods imply. Terms are used without asking whether they represent something that really exists in the extrasubjective* world or without knowledge of a detailed description of the process named with a term (e.g., "principle of reciprocal illumination" or "long-branch-problem", ch. 5.1.1, 6.3.2). The considerations in the following paragraphs are meant to call to mind that science and human perception are always entailed with the use of language. Be encouraged to always ask for the elementary roots of hypotheses and knowledge.

1.1 What is knowledge?

Knowledge is the result of a cognitive act. Knowledge about facts is not gained through thinking in the first place: primarily we have to perceive the world. Cognition is the **perception of existing facts**. It is an achievement of our nervous system in collaboration with our sense organs, and consists of the recognition of constant patterns or of patterns that are repeatedly seen and used to identify and classify things, sound or smells (properties of things) among a large diversity of sensations (see Oeser 1976). Even the simplest perception requires this achievement. There is no doubt that even a dog can identify existing facts. This "pure recognition" or "perception" (subjective experience) exists independently of language. We react quickly to dangers without having to express the recognition of danger in words. Human perception also implies the formation of notions not based on language. This also occurs in a simple form in animals: a dog undoubtedly knows bones, traffic-lights and stones. Monkeys can associate notions and symbols and use them in an experiment to communicate with the researcher. However, to transform this perception into transmittable knowledge we have to use language, a characteristic feature of human cognition. Only experiences and findings which can be passed on through language are intersubjectively testable. Therefore, often only the notional and linguistically fixed human knowledge is called knowledge in the strict sense.

Scientific knowledge on the other hand is the result of critically tested perception, the directed search for falsifiable theories to explain observations of nature (Popper 1934), and not only a collection of experiences and subjective convictions based on observations, nor is it only the result of the application of the rules of logic (see ch. 1.4.7). Hypotheses or proposed explanations

* *extrasubjective*: something outside the mind of a thinking subject.

have to serve as a basis to decide which experiences and observations shall be collected in nature or in an experiment. These will be used in turn to test whether the hypothesis proves successful or not. Also, one's own perception is tested in this way. In each branch of science, specific methods have been developed and the experts of each branch decide which methods are accepted as being "scientific". A universally valid criterion to classify methods according to their scientific value does not exist, but there are the statements of experts who claim that they were able to gain new knowledge with specific methods.

Epistemology: methodical reconstruction of the process that leads to the acquisition of knowledge.

Cognition: perception of existing facts, a process whose outcome is knowledge.

Logic: collection of laws used to interlink statements in order to deduce valid conclusions.

Theory of science: expansion and use of epistemology and logic for the analysis of scientific cognition.

Scientific cognition: systematically planned acquisition of knowledge.

1.2 Classification and the function of language

The conscious use of language is essential for the communication of scientists and for the development of scientific theories. Numerous misunderstandings have arisen from the confusion of theoretical concepts and terms with extrasubjective reality. For a deeper understanding of this subject I recommend reading scientific publications on the theory of science (e.g., Oeser 1976, Seiffert 1991, Janich 1997, Mahner & Bunge 1997).

In our world there are **objects** we can name using words thanks to our genetically based faculty of speech. We learn these words in a pre-scientific phase when objects are shown and explained to a child or some other learning person. An object can be a tangible natural thing, but also a technical instrument, or a process or property, or any subject one is talking about. Therefore, an object in our language does not necessarily have to be a "material thing per se". The ontological status, the question of the real being of the object does not matter. We can talk about something we have experienced or even invented. Some language analysts call words that name objects "predicators". **Predicators** can represent material objects (e.g., a "chair"), but also properties of these objects ("wooden"). Strictly speaking **we cannot differentiate between the properties and the object itself**: we always perceive some properties but not the "being as such". A "chair" is a construction with a back and it can serve as a seat for a person. We identify a thing as "chair" because we recognize its function and specific properties. The thing "wooden chair" is a subset of the set "chair". It has to be simultaneously "chair" as well as "wood". In this context the predicators "wooden" and "chair" are logically of equal quality and interchangeable: "the chair made of wood" and "the wood in form of a chair" include the same set of objects. The naming depends on our subjective intention: the "wood for burning" can be a chair or have other shapes, only the material is important for the user. The "chair for sitting" can be made of wood or some other material, what is important is the function.

The predicator "cow" at first is nothing more than a word we can use to refer to specific properties of a thing. Whether the object "really exists", is present "per se" (existing "extrasubjectively") is not relevant for the development of our vocabulary. In natural sciences, however, it is essential to find out whether these predicators refer to real objects, to properties of things or processes existing outside of our minds, or if they are a product of our thinking and our errors. It has to be investigated which properties of all the objects we call "cows" are held in common, and whether people using the word "cow" refer to these properties.

Obviously a predicator is used to refer to a number of things which share common properties. We sometimes get overlapping sets ("wood" and "chair"), but there are also **hierarchies**. That means that a predicator names a subset of another predicator. Examples are "(vehicle (motor vehicle (car (all-terrain vehicle))))", also the hierarchy "quad-

rupeds (mammals (ungulates (cow)))" is not logically different. This hierarchy does not state anything about the origin of these things or about genealogical relationships existing between them. Each ungulate possesses specific properties also occurring in a cow, but not all of them. Each mammal has some of the specific properties found also in ungulates, but not all of them.

The ability to classify things according to their properties and the potential to refer to each of these combinations of properties with a word is **inborn** to humans. We build hierarchies of terms on the basis of common properties without the need for substantiation of the material existence of these hierarchies. The hierarchies result from our innate ability to classify things.

This shows that classification is a pre-scientific activity (see also ch. 1.1). The nested terms, the encaptically arranged (hierarchical) notions are **classes**: a class is an intellectual concept, enclosing a number of things that share the same properties. The word we use for this is a predicator.

Thus **classes** are notions which can be used to name natural assemblages of objects but also artificial groupings. "The chair per se" does not exist. Chairs differ in origin, material, form, and colour. The term "chair" represents a class, as does the term "horse". The latter names a grouping of objects whose common characteristics were determined by natural processes (a natural class or "natural kind": Mahner & Bunge 1997). Numerous names of animals are artificial groupings: for example, in English aquatic animals are called "fish" disregarding their construction, lifestyle or phylogenetic relationships: jellyfish, crayfish, starfish, shellfish. Similarly, a "worm" can be an insect larva, an oligochaete, or a nematode.

To differentiate hierarchical levels, i.e. different degrees of abstraction, we could introduce **categories** or "ranks" for each level. A "special type of honey" would be a subset of all sorts of honey (= honey species) with specific properties. On a higher level the term "genus" or "class" could be used. In colloquial language however, these terms (species, sort, genus, class) are used synonymously. This indicates that these categories are chosen subjectively and that there are no measurable or tangible properties which would enable a differentiation of these categories. "A queer fish" is also "a special sort of human", "belongs to a rare genus of human", and is "a class of its own". These categories only express that a subset of "man" is concerned.

If we assign a **proper name,** it concerns a single historical object, an individual of limited existence in time. It can be a real material object or a mental construction ("Donald Duck"). Proper names are used without the necessity to assign a predicator to the individual: we can talk about Charles Darwin without having to associate the name with a biological species affinity, a profession or nationality. Therefore proper names cannot be used for other individuals "of the same sort" and are not appropriate to name universal processes or abundant things. The proper name does not give away anything about the properties of the individual, it does not belong to the common vocabulary of our language.

There are also **hierarchies of proper names**: the individual "Germany" includes the individuals "Hamburg" and "Berlin"; the street "Reeperbahn" belongs to "Hamburg" and "Unter den Linden" belongs to "Berlin". A corresponding hierarchy of objects (more precisely: an accumulation of material things) does really exist, but it cannot be *described* with a proper name. Whereas for a yet unknown object (e.g., an unknown plant) predicators can be assigned according to its properties (e.g., bush with rose-like flowers) and the object can therefore be fitted into a hierarchy of predicators, this is not possible for proper names. Seeing an unknown town it will not be apparent that it is called "Berlin" and is situated in Central Europe, it could also be found in North America. It is impossible to recognize the individual town without knowledge of the peculiarities of "Berlin in Germany". These peculiarities are not entailed with the word "Berlin", as shown by the fact that a village in another part of Germany is also called "Berlin". These considerations become important when we reflect upon the utility and the ontological status of names of groups of organisms (see ch. 3.1 and 3.5).

A **classification** can only be realized with predicators, not with proper names. The predicators define the properties of the members of a class. This seems to be in contradiction to the fact that groups of organisms get proper names ("Equidae"). However, the same group *as a class* is de-

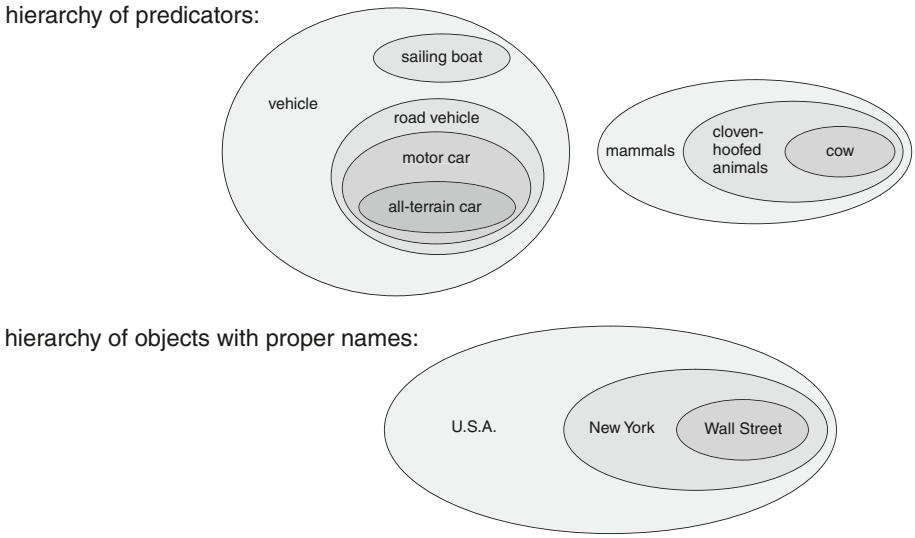


Fig. 1. Hierarchies.

fined with the predicator "horse-like", while the *mental object*, the group of all horse-like animals, gets a proper name. The hierarchy of proper names refers to an "objective hierarchy", when single material or mental objects are real parts of a larger whole. We do not have to know and name the properties, the hierarchy exists anyway. It is different in the case of a classification: the objects (e.g., single animals) are independent of each other, we classify them according to aspects of convenience for communication, and we mentally associate function or property with a word (predicator).

A hierarchy, which at the same time is a relationship of sets, can be described with a Venn-diagram (Fig. 1).

This example elucidates that we do not have to discuss the "real existence", the ontological status, of the object "ungulate" or "car" as a material object or process or system in nature, as long as everybody knows which properties are meant by the word. A prerequisite for communication is that every interlocutor associates the word with the same properties; he does not have to recall a real existing ungulate. It is reasonable to assign a single word to the property "animal with paired hooves", because we encounter this "sort" of animal often in everyday life and we do not always know the appropriate name (e.g., antelope).

It is different with proper names: we can only talk about "Hamburg" if the interlocutor knows Hamburg or when he is aware of the fact that this is a specific, individual city.

The question may be asked whether "Mammalia" is a predicator or a proper name. Used as predicator it refers to suckling as a character of mammals, used as proper name the word is assigned to an individual object. Considering the fact that there are also mammals (Monotremata) which do not breast-feed their young and do not have nipples, the status as predicator has to be denied. If the word is a proper name there has to be a corresponding individual object. It can be shown that this object is a conceptional individual* and thus a construct. (The terms relevant in this context are discussed in later chapters: monophyla: ch. 2.6; taxon: ch. 3.5; biological classification: ch. 12.).

The **pre-scientific classification of living organisms** is primarily a practical one. Practicality in this respect is determined by human cognitive and communicative abilities. This classification does not state anything about the real genealog-

* *conceptional individual*: an individual that exists only in our minds.

ical relationships of living organisms. The logical relationships are the same as in the classification of "vehicles".

Assigning a term to a thing implies a certain amount of information about the thing as the term refers to invariable properties occurring in different things. With this association (classification) a statement about the properties of the object is implied. The number of properties associated with a term varies greatly; general terms (vehicle, mammal) imply fewer properties common to all included objects and thus less information than specific terms (bike, cow).

As we learn to designate things with language, terms are initially not defined. A **definition** is "equalling a yet unknown word with a combination of at least two known words" (Seiffert 1991): a "pinto" is a "piebald horse", thus the intersecting set of "piebald" and "horse". To define a new word, other words have to be available. In order to find out whether the new word and its definition name a thing of the extrasubjective reality, each word of the definition has to be checked. As these words can also be defined by other ones, we have to search for the origin of these words in colloquial language. There we can encounter "point zero", the undefined words of pre-scientific language. The child is introduced to words through reference to examples ("this is a car; this is also a car") and not primarily through definition, because definitions have to refer back to known words. Thanks to its powerful data processing system (nervous system), the child has the ability to identify the common attributes of cars.

As there may be several words for the same object (synonyms: horse, nag, steed), obviously besides these words there exists an abstraction not directly interlinked with a specific term, although the idea of what the abstraction reflects developed through the use of the words. This abstraction is a **notion**, which exists independently of the type of spoken language and can be named with different words. Through examples and practical usage we learn which objects, characteristics or processes in nature correspond to a notion and which words are available for it (see ch. 1.1).

A notion is always represented by one or more words, but not necessarily by only one specific word. Therefore, a notion cannot be defined: the words "iron" and "element with the ordinal number 26 and the relative atomic mass 55.8" represent the same chemical element, they represent the same notion. A notion of this substance develops through experience but not through definition.

Object: something we can talk about (a material object, an observation, an idea, etc ....).

Thing: material (existing) object.

Fact: a material object or state of an object, or an event occurring in the material world.

Construct: object not existing outside the thinking subject (conceptional object).

Predicator: a word naming a thing.

Notion: something we can refer to using synonymous words (e.g., abstraction for the common meaning of horse, nag, steed). Words can be defined, notions cannot.

Definition: equaling an unknown word with a combination of at least two known ones.

Term: a predicator used in science and technology which is introduced through definition (explicitly) and/or with examples (exemplary). Its meaning is determined through convention.

Class: construct for a group of objects that share a certain property (term "classification": see ch. 1.3.4; term "category": see ch. 3.5).

Material individual: single, material, physically limited object, existing independently from the observer.

Conceptional individual: a construct, a mental individual (e.g., Donald Duck, Hamburg).

The analysis of the language-reality relationship could stop at the point where we discover the origin of words in colloquial language. This, however, would be a source of uncertainty: does the word "red" name something perceived in the same way by each human because it is reality in the world outside our individual perception, or does each observer use the same word for different colours or properties? What does really exist in nature?

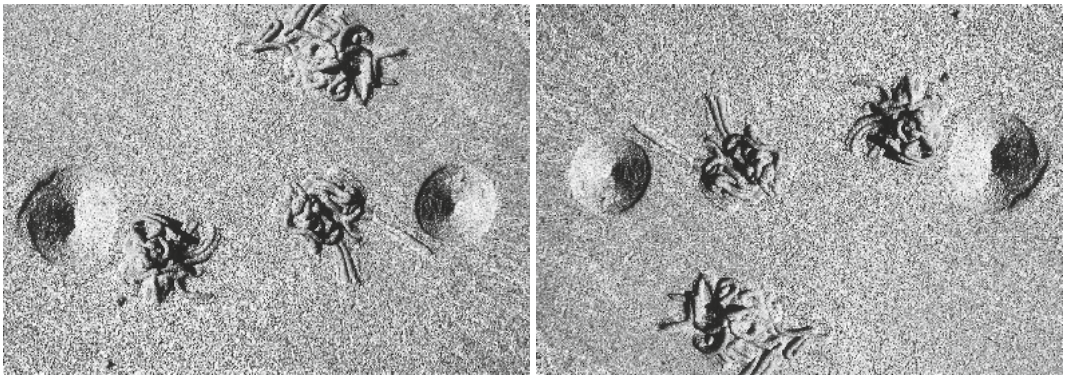


Fig. 2. Optical illusion: funnel in the sand or cone of sand? The interpretation of both photos is very different, although it is the same picture, one turned 180°. The interpretation is based on our experiences with shadows of three-dimensional structures and is an achievement of the central nervous system. (Feeding funnel and excrements of *Arenicola marina* (Polychaeta)).

At this point we realize that our brain gets its whole "knowledge" of the surrounding reality through sensors, which forward signals about our environment to the brain. Whatever our sensory organs transmit to the data processing organs determines our view of shapes and properties of our environment. Can we trust our senses? Obviously our "sensors" do not transmit everything that can be registered and often even wrong information: we do not sense magnetic fields (at least not consciously), we do not see UV radiation, a blow on the eye causes the impression of light although there is no light. Additionally, our brain interprets the sensory impressions in a way that is not consciously controllable. Already, at the unconscious level of processing of information in the central nervous system, signals transmitted from sensory organs are evaluated according to their fit to inborn or learned patterns: when we put on glasses turning an image 180°, we adapt to this after a while and see the picture "in the right way". Because we are used to shadows produced by light shining from above, we interpret silhouettes accordingly (see Fig. 2).

Thus there is reason to check which of the objects we name really exist outside our minds. Ultimately we have to analyse the root of our uncertainty: can we trust our "cognitive apparatus" and thus our pre-scientific images of the world? Will inborn processes alter the incoming data in such a way, that they do not allow a conception close to reality? An answer to these questions is given by evolutionary epistemology (ch.1.5). But first we will consider some aspects apart from our cognition apparatus.

1.3 What is there outside of our cognition apparatus? What is "really existing"?

Not all objects of conversations exist outside the subject, and not all words referring to material objects name individual things. We have to distinguish carefully between facts existing in nature and constructs of our thinking. Therefore the following examples are discussed so that everybody can check for themselves whether she or he is aware of the differences.

Everybody knows what a "wood" is. However, is there a real entity in nature, existing as a "wood" independently from our cognition? Looking around a landscape we find trees and assemblages of trees. Comparing groups of trees we can find a continuum of increasing numbers of trees per hectare (increasing abundance), and of the size of tree-covered areas. A natural graduation of classes of groups of trees, which could be iden-

tified objectively, does not exist. Thus the word "wood" means a subjectively chosen "large number of trees standing close to each other". Depending on the person they may be experienced as a plantation or as a wood of fairy tales. A specific wood would be an "entity of nature" if it could be proven that all trees therein depend on each other, are linked with each other through processes or other relationships, and so form an individual material system with typical peculiarities. Undoubtedly, a mutual relation between trees exists, e.g., through shadowing or competition of the roots. These influences however exist with each neighbouring plant of a tree, whether the tree grows in a "wood" or in a "park". The occurring processes are not a specific property of a system "wood". An attempt to find the "wood per se" in nature will not be successful. The word "wood" is associated with specific experiences we have had, and these can be summarized economically with the word.

The word "tree" has the same ontological status: it is a predicator referring to well-known properties (large plant with a wooden trunk). The "tree per se" does not exist, the notion comprises very different sorts of plants. Additionally, the same individual plant may look like a "shrub" or "bush" while young, and only becomes a "tree" years later: its properties and therewith the appropriate predicators have changed. Let us refine this example: an oak tree is a subgroup of "tree". The word "oak tree" is also a predicator used for objects with specific properties. A special oak tree, planted by the President of the German Federal Republic on the 25.10.1966 in the city park of Schildburg, is an individual. This tree could have a proper name, can be perceived as whole and does exist for our sensory organs as a defined object. We can identify it by the coordinates of its location and the "oak tree specific properties".

We notice however, that the appearance of the individual changes with time. Obviously, it is not the individual parts or components (specific branches, vessels, cells, individual ribosomes) that render a tree an individual, but the cohesion of all parts of a single material object that can be detached from its surroundings at a given time. The material components and appearance of the tree change continuously through processes that link the components of the individual to a functional whole. The components are dependent on each other (mechanical support, supply of water and nutrients). Thus this oak tree also is a **material system** (see ch. 1.3.2). The system develops during the course of time and all components take part in this development. The death of the system marks the end of all mutually dependent components (living cells in leaves, root, trunk, …); only the parts which can exist without the system (pieces of wood, water molecules, ions) remain. The system is "open", matter and energy flow through the system, matter and energy-rich molecules are emitted from the system. Despite this continuous change, the system exists as an individual historical entity.

1.3.1 Objects of nature, the "thing per se"

In the following we will only talk about things that exist as real material entities outside of our minds. Individual, physically delimited objects are called "things" in colloquial language.

A modern scientist is aware of the fact that **things** in nature with great probability have their own individual history that, according to our present view, started with the phase of the cosmic big bang. Each individual stone, a specific planet or tree are objects which developed historically. They have a defined time of existence, which in some cases may be long. Each thing thus has the characteristics of an individual: it comes into being, it exists, it vanishes. Individuals can be given proper names as done for example for pets, planets, precious diamonds or buildings. The lack of proper names does not change the fact that such things are individuals.

Material objects consist of components of which the smallest entities are studied by nuclear physicists. When a stone decays to sand, usually the chemical composition of the crumbling material is still the same. What has changed? What was the property that made the stone a single "thing"? Obviously we recognize objects as single things only if they can be moved independently of their surroundings as a unit containing all atomic components. Apparently a tree cannot be moved, but it is possible if we unearth it; we also experience that storms may uproot trees. The roots are not fused with the soil but with the tree. A rock immersed and fused with the mountain cannot

be identified as individual rock. When blown out with dynamite it gets a life of its own with this "hour of birth", it can develop and move independently from the mountain. A "thing" or "object in nature" is nothing else but a notion for matter kept together by physical forces and detached from the surroundings. A sugar crystal is a thing which gets smaller as soon as it is placed in a glass of water and finally disappears, although the same amount of sugar is still present in the glass. A large thing can be partitioned into smaller things.

The notion "thing" serves to classify the phenomena of nature described above. The individual thing exists outside of our consciousness, such as the 109-carat diamond of the British crown-jewels called "Kohinoor".

Now we have to ask whether the Nile, the Atlantic Ocean, Mt. Vesuvius are individual things. Obviously these objects cannot be separated physically from their surroundings. Their existence depends on the surroundings, providing water or lava and ashes, there are no material borders between the named structure and the surroundings. Nevertheless we can recognize them as individuals and name them: there exist **material systems in nature** or subjectively delimited parts of larger systems.

Groups of objects: an accumulation of objects often is explicitly named (dune, forest, town, herd). This mental grouping may induce one to consider these accumulations to be real things. Reality is only the existence of individual delimited things, their special closeness as well as any processes occurring between them. When the existence of the group or its properties is dependant on its own processes, then the group is a material system or part of a system.

1.3.2 Systems

Before we can discuss the ontology of a "system of the animal kingdom", we consider the notion "system" more generally starting with its use in everyday colloquial language.

A system is a grouping of things or statements, linked by relationships that integrate all parts to a more complex whole. We can identify the system because it reacts as a single unit. A system can be a principle of order invented by humans and existing only in our minds, such as a grouping of notions, statements or perceptions (**intellectual system**). A system can also be something existing in reality outside of our thinking (**material system**) (Seiffert 1991). Natural scientists are primarily dedicated to the analysis of material systems. These can be created by humans or they can exist in nature without human interference.

Examples of systems created by humans are: a library, a calculator, a radio, a state, a staff, a system of notation, a dictionary. Even though some of these systems are material objects, they are also based on intellectual systems used as analogous models for the material thing.

Examples of systems in nature: a river, a fire, a coral reef, our planetary system, a tree, a blood circulation system.

In each case parts of the material system are interchangeable: single (not all) books stored in a library, words in a dictionary, wires in a calculating machine, a colony of animals of a reef, single cells or leaves of a tree can be taken out or replaced without stopping the existence of the system. A river exists because water is continuously supplied and – following gravity – is directed in specific courses due to the topology and permeability of the ground. This is a system with blurred borders, because water can seep away laterally at the banks, or some water can run off underground. The river we call "the Nile" exists without doubt, despite its fuzzy delimitation, just as a certain cloud can exist for a limited amount of time. The same is true for a blood circulation system which is partially open: the fluid can also circulate between tissues outside the blood vessels, a strict boundary between "inside" and "outside" does not exist for this system.

As long as we look at these systems at the present time, we can often clearly separate them from each other. However, problems occur when we look at the development of the systems **along the time axis**. Where do systems start, where do they end? The beginning is obviously to be found where the process joining the single parts of the system starts. Historically a river starts with the first runoff cutting its path to the ocean, deepening its bed. A fire starts with the process of igni-

tion. But where is the beginning of a tree? In the seed? Or in the seed of the parents? And is a solar system whose sun has only "captured" five planets a totally different one from a system with the same sun but with seven planets many millions of years later?

In the sense of colloquial language, systems are not required to have a sharp boundary along the "time axis", but in a "time horizon", a point in the dimension of time. Fire may serve as an example: a fire is more than just a chemical process. Material things such as gases, wood, products of combustion are parts of the system. The individual system only exists through the process of combustion. When a fire divides in a savannah and two fires continue to burn, two new independent systems have originated. The moment of transition cannot be determined precisely, and thus the end of the first system cannot be exactly ascertained. Viewed objectively, the process itself is not interrupted but the things involved are replaced. Thus a delimited material system really exists only in a given time horizon. The mental connection of objects and processes with the past serves to reconstruct the course of a process, and to understand the exchange of the elements involved.

Each reproductive community of organisms is a system comparable to the fire described above. The "process of life", of which reproduction is a part, has taken place continuously since the occurrence of the first living cell. The things involved, however, are "used up" like the wood in the fire and are replaced by new ones, and entities originate which lose physical and functional contact with each other (see divergence of populations: ch. 2.2).

Systems existing in nature have the character of individuals: they originate, exist as single units for some time, and fall to pieces again. Systems sharing common characteristics can be named with a predicator ("river"). However, the "river per se" is not really existing, but only the instances we call "Amazon" or "Nile". With the notion "river" we name what we know as common characteristics of all real rivers. Our notion "river" is a mental model containing the common properties of real river systems we know.

For the existence of a material system as an individual it is important that all its parts have certain properties and that the relationships between them continue. In material systems these relationships are physical forces or processes influencing the fate of the individual parts (the water molecule in a river is pushed, the cells in a tree are nourished, the book in the library is shelved to "its" location). Each system is an individual entity with its own history. Libraries can be organized according to different principles. Their individual properties are determined by the principles of order that represent the relationships between the parts, and by the properties of the parts of the system (books, catalogues, librarians). The fate, the longevity of the system, depends on these properties. The system differs from the sum of its parts, because within the system the individual parts have a common fate through mutual influences, but develop independently outside the system.

These examples illustrate that systems do not necessarily have to be units with a strict construction; especially living organisms are open, variable systems. All systems are in contact with their surroundings, they are never absolutely "closed". Natural systems are as real as the processes and the material things within them, their delimitation from the surroundings is often an abstraction. With their separation along the time axis we create boundaries that do not exist in nature, and thus our concepts of such "parts of a system" are constructs (see "biological species" ch. 2.3). Descriptions of systems in four dimensions are mental images.

1.3.3 Thing and system

Not every system is a thing (see above). Each thing consisting of more than one elementary particle is a system: on first sight in a crystal no processes take place which influence or change its component parts. At a second view however, one can detect processes such as the change of the crystallite structure and the oxidation of elements. There are forces binding the atoms in a certain order. The crystal can grow and catch molecules from the surroundings, atoms or molecules are "integrated into the order". Such systems consisting of atoms or molecules have new properties (e.g., refraction) not shown by the isolated components.

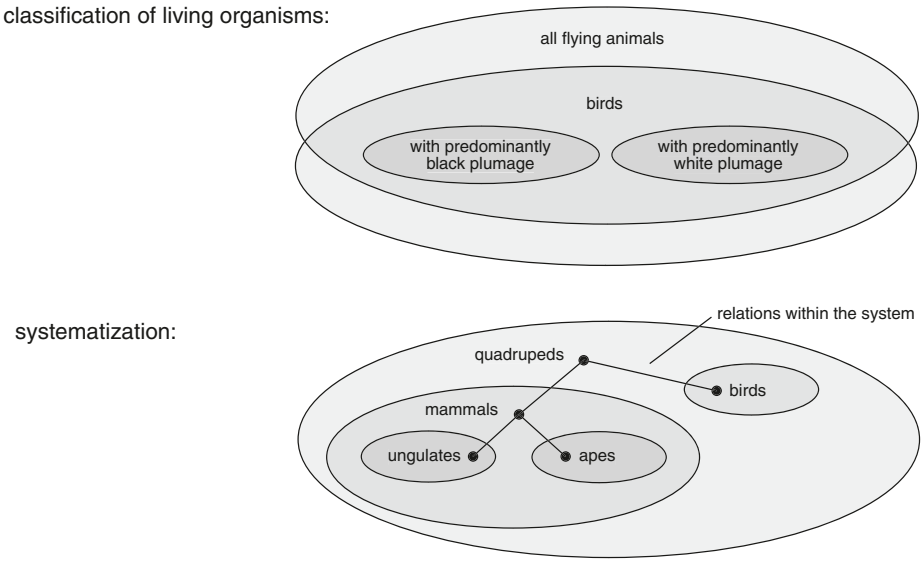


Fig. 3. Classification and systematization: each classification of organisms can, but does not have to, contain monophyletic groups (term monophyly: see ch. 2.6). Within phylogenetic systems the inherent relationships are relationships of the mental system, namely the hypothetical phylogenetic relationships. The hypotheses are depicted in the form of a tree graph. In the graph the nodes (= vertices) represent species that are part of higher ranking groups, the lines represent ancestor-descendant relationships.

Material object: material things linked by physical forces, forming a unit which can be moved independently from the surroundings.

System: a set of intellectual or material objects which lawfully interact with each other.

1.3.4 What is a "system in the animal kingdom"?

We have to differentiate the two basic types of systems: mental and material systems. A set of predicators forming a hierarchical system in the sense of chapter 1.2 first of all is a mental system. Whether this set concerns organisms or vehicles is irrelevant (Fig. 1). The system of organisms in the sense of Carl von Linné (1707–1778) is an intellectual system, originally designed to classify the (bio)diversity according to arbitrary rules (see "classification" ch. 1.2; "categories" ch. 3.7). Here one criterion for classification is the **perceived similarity**: Linné orders flowering plants according to the number and position of styles and stamens. A group of species which has been intellectually united and named is a **taxon** (ch. 3.5). Interestingly, Linné considered his higher taxonomic units to be artificial groupings although he intended to record the "natural order" of the creation plan. The phylogenetic system on the other hand refers to **historical processes** which have to be reconstructed. Note, the classification of organisms is not the hierarchical set of proper names given to groups of organisms but a mental hierarchy of properties generally attributed to groups of organisms: snakes and birds are called vertebrates, because they have a spinal column.

It is often stated that we are in search of the **objective system**, the "order of organisms" existing in nature. This "**natural system**" should still exist even when humans vanish. But viewed realistically, groups of organisms are not linked by system processes correlated with genealogy. There are ecological effects in a given time horizon ("now"), however, all mammals living today do not form a functional unit that also develops as a unit. A real material system "mammals" does not exist. Regular interactions between mammals such as a fox in Scandinavia and a bear in North America do not exist today. The "order of organisms" is a mental order. What exists in reality are the properties of individual organisms which can be similar in different individuals.

What the systematist looks for is the reconstruction of historical processes, the sequence of speciation events (explanation of the term speciation in ch. 2.3.1). This sequence is depicted as a "phylogenetic tree", which is a construct. The entities of descent are called monophyla (see chapter 2.6) and in many cases are referred to with proper names. The theory-dependent representation of this sequence of events is called a "**phylogenetic system**". This is not a material system of organisms. As the phylogenetic system is a representation based on a reconstruction it is always hypothetical, and it is always a construct. In the same way the "order in nature" which is sought by systematists is only a useful metaphor. Order exists in natural history collections (where all butterflies can be found on the same shelves) or in compatible hypotheses of monophyly.

The search for a **phylogenetic system** of organisms proved to be necessary because the classification, which serves to master the diversity of life, can only be tested objectively with reference to inferred historical events. In biology we call this search for phylogenetic relationships **systematization** to differentiate it from **classification** (Ax 1987, 1988).

A pre-scientific classification of objects leads to an artificial diagnostic and organizing scheme. As long as the classification is not the representation of a theory, it is only based on conventions for the use of predicators. However, even when succeeding a scientific analysis, a classification depends on *predicators*, i.e. some properties shared by all members of a single class. The biological systemization, however, is not based primarily on visible features but on hypotheses of phylogenetic relationships. It is a hierarchical grouping of monophyla following rules based on a theory and thus has a higher information content. It is not the convenience of terms for the daily use of laymen that determines whether a concept for the delimitation of a taxon is accepted, but the usefulness for the communication of scientists. A consequence of the systematization is the classification of properties of organisms:

Vertebrata are animals with bones; the monophylum Aves is defined as Vertebrata with wings and feathers; this in turn includes the monophylum Spheniscidae (penguins), defined as Aves with short, stiff, finlike wings. The system of organisms is not only a hierarchical set of anatomical details and of visible similarities, but of all biological properties including the relationships to the environment. This is the reason why **the explanatory and prognostic power** of the phylogenetic system is much higher than that of a different artificial system. Historical biology has to deal with processes which gave rise to the organisms, but it is not the history of mental systems.

Classification: construction of a mental system of *predicators*, the order of which is determined by the usefulness for verbal communication.

Systematization: Hierarchical grouping of *proper names* of monophyletic groups of organisms into a mental system representing an inferred sequence of speciation events (term speciation: ch. 2.3.1; further details in ch. 12).

Phylogenetic system: notional or graphic representation of the mental hierarchical order of taxa which can be deduced from the systematization of organisms.

1.3.5 What is "information"?

Information theory cannot be presented here in detail (see Shannon 1948, Hassenstein 1966, Oeser 1976, Wiley 1988, Cover & Thomas 1991, Schneider 1996, Mahner & Bunge 1997 for further details). However, the question of what the notion "information" denotes needs further consideration, because you will encounter it frequently in the following chapters. The following list contains contexts in which the word "information" is used:

- Spoken words, written words, traffic symbols transmit information. An architect gains information from a construction plan.
- Radio stations transmit information using electromagnetic waves.
- A witness informs his audience.
- A computer processes information.
- The DNA of an organism contains information, which plays a central role for its construction during ontogenesis.
- A pheromone transmits information concerning the presence of sexually mature mates of the same species.
- Homologue characters contain information about properties of ancestors.

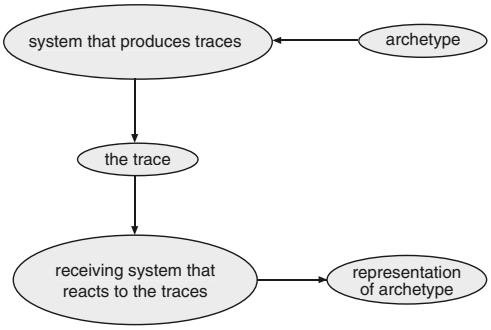


Fig. 4. Process of "transmission of information". Technically, the "system" creating traces consists of a "coding unit" and the "transmitter". The "trace" is the signal existing in a mediating medium or moving in a "channel". The "receiving system" consists of the "receiver" and the "decoder". Decoding is the transformation of the information to a specific reaction of the receiver.

Obviously there exist **carriers of information** such as symbols, drawings, sound waves, or macromolecules which can cause something. They can only affect **receivers** adapted to specific types of symbols or to which the symbols fit to: the blueprint fits to the trained architect; the DNA fits to the cellular apparatus of an organism; the electromagnetic signals are compatible with a specific computer system; the words can only be understood by someone knowing the language. Data are represented without the need to move their source, the real objects, around. A drawing in a fashion journal can trigger the sewing of dresses. Fashionable details like the cleft between collar and lapel in jackets and dresses are transmitted from generation to generation like biological heritable information. Whereas in the time of Friedrich the Great (King of Prussia) decorative embroidery bordered the buttonholes of officers' jackets, later these patterns were detached from the original purpose and used as signs of honour on collar plates. The homology of the pattern is obvious (see Koenig 1975). In the sense of colloquial language, symbols with information are an abstract representation of a fact or an object, they "represent something" (but this is not always so; see below).

Information influences the receiver by inducing processes: learning processes, electrical currents in technical devices, movements, behaviour. If no appropriate receiver exists, the information (in the sense used in colloquial language) is "worthless". Information may be tracks left by some process. In this sense, a book contains information in the form of traces of a series of mental processes. The content of a book only becomes "knowledge" when it is read *and* understood. The scribbling of children does not contain information in the sense of description of facts. However, a psychologist or educationist can draw information on the state of development of the child from the scribbling: in this case it is obvious that information is more than only "news", namely a trace of human activity. The recipient of information in this case would be the trained education specialist who can analyse this kind of information (e.g., indications of the abstraction capabilities of the child and of its manual skills). We can abstract further: each **process leaving traces in nature**, leaves **evidence of one or several properties of these processes**. This can also be taken literally (tracks of dinosaurs). Whether a pattern we notice in nature contains information or not depends on whether we are able to identify in it evidence that allows conclusions about the processes which caused these traces.

It proved useful to use the notion "information" only in cases where the existence of an appropriate receiver is assumed: information consists of symbols or traces which transmit something about a specific "source" to a **receiver**, or which induce a specific behaviour of the recipient depending on the information (e.g., foraging). A recipient which is adapted or trained to read these traces has to be present. In this context to "read" means that specific processes are induced in the receiving system depending on the type of traces. If there were not a system in the universe which is able to deduce the existence of dinosaurs from the existence of dinosaur tracks and bones, then any knowledge of dinosaurs would be lacking. In nature, first of all, there exists the material trace. It contains "information" only for someone who can read it. The material trace (or the symbol, the sign, the series of words) is not an abstraction. An abstraction is the linguistic notion for the whole process.

From the point of view of physics the transmission of information is always coupled with the transmission of energy and entropy, whereby the energy transmission is irrelevant: the smallest amounts of light can be sufficient to allow reading of a serious message. Entropy as

the measure of "disorder" decreases in a system through the gain of information. Phrased differently, the gain of information of a system means a local increase of order. Since entropy cannot be destroyed, the system has to be open and has to emit entropy to the environment when it gains information (see Ebeling 1990). The emission of entropy can for example take place through radiation: the surroundings warm up. The notion of information presented here implies that the physical signal (the "trace") is not the information but rather what the recipient reads from the signal or what is coded in the signal (e.g., instruction for an action).

We refer to information even if it "rests" (e.g., in libraries), in cases when a potential recipient is present but not active. If there are many traces, a lot of information is present. Depending on the decoder of the receiver, a large trace could convey a small amount of information, or vice versa. With this statement it becomes clear that we can describe this process quantitatively (see ch. 5.1).

In everyday life there are cases where originals and copies exist. Copies and originals can be material or mental objects. For the newsreader of a radio station the written sentence is the original, for a painter the real landscape, for the architect the idea existing in his brain. In the case of dinosaur tracks, the coding system that produces traces is identical with the original. A thought of a speaker will develop to a similar "appearance" in the minds of a listener. A house can be the exact copy of an original if constructed according to the same construction plan. But the idea for the first house developed without an equivalent material original.

We are used to associating the term "information" with the idea that signals or words "represent" something: a photo represents a real object, a report represents a real event. Such information has a **meaning** that can be subjectively evaluated. This, however, is only one special case of information transmission. The DNA for example is not a "copy" of an organism in the sense of a representation of the construction that is readable for humans. Rather the DNA contains coded "instructions" which markedly influence the construction of an organism through complex processes. When a robot, whose task it is to weld two pieces of metal together, is fed with information, a replica of the original does not develop in the robot, rather a reaction typical for this system will follow. For the robot this information has no "meaning".

In biology "meaning" can be attributed to DNA molecules only in the sense that genes are correlates of proteins and control mechanisms, which, amongst other things, contribute to the construction of cells and organisms. The process whose traces the genes are is evolution; mutations which do not fit to the cell apparatus or to the environment of the organism are not retained within a population. Copies of more suitable genes are passed on to further generations in their place. In this way the DNA of an organism contains traces of the physical and biological environment (Goethe (German poet) wrote in 1824: "if the eye weren't sun-like, it could never see the sun"). The selecting forces of the environment serve as the "original". During the course of evolution "**genetic information**" accumulates in organisms, a phenomenon called **anagenesis**. The receiver can be the cellular apparatus, but also the geneticist searching for different information than the systematist; the quality of the information is a different one in each case. The systematists do not have to know the functional receiver (the cellular apparatus) of the information, they rather analyse the process of information transmission from ancestors to descendants. Thus they do not analyse the "meaning of the radio show" to the listener, but the quality of the transmission from sender to receiver and the identity of the sender (see definition of homology: ch. 5.1)

We will call those traces which were left by a process and that are readable for a receiver "**signal**". Signals can be material things in a certain state or processes (e.g., production of sound waves).

The process of information transmission implies the phenomenon that signals or traces can become blurred. This "**noise**" has the effect that depending on the degree of signal destruction the reaction of the receiver is not the same as with an unaltered signal. In the worst case the reaction is wanting: the information has been lost.

The term "**noise**" can have two meanings: it can be the process changing the signal, or the result of this process. The latter is the meaning relevant for phylogenetics.

Information: the fact that a trace (also symbols, strings, vibrations), that was produced by a process or by a transmitting system, can influence a reaction of a special receiving system in a specific way. In this case the trace is "informative".

Signal: a different word for the notion "informative trace".

Noise: a) process modifying signals during transmission, but also b) signals which have been modified during transmission. Noisy signals can induce a different reaction in the receiving system than the unchanged signal would.

Genetic information: the fact that a specific structure of DNA-molecules triggers a specific process in a particular receiving system (cell, in vitro system, cognition apparatus of a scientist).

Anagenesis: Accumulation of genetic information with time.

1.3.6 Quantifying information

The notion "information content" implies the necessity to quantify information. The information content can only be estimated in context with a specific transmitter-receiver system, whereby the reaction of the receiver has to be used as standard. Intuitively it is clear that the precision of a statement in a sentence increases the more it excludes. The statement "a distance of about 100 km" can mean 88, 100, or more kilometres, while the specification "106.5 km" excludes the given alternatives and is therefore more informative. The quantity of information increases through addition of single pieces of information: a dictionary with 20 volumes contains more information than a dictionary consisting of only one volume. To understand how the amount of information can be described objectively, it is worthwhile to have a quick look at the concept used in computer science.

Shannon (1948) established mathematical communication theory, which analyses the statistical basis for the transmission of information, in the simplest case based on binary coding. Shannon's concept serves in general to measure the minimum complexity of patterns necessary to code a specific information. **Shannon's information concept** (1948) is adapted to the quantitative description of transmission errors. Hereby the "meaning", that is the quality of the reaction in the receiver, is of no importance. It is assumed that the receiver has constant properties and can read the incoming symbols.

The notion "information" in computer science is interlinked with an aspect of uncertainty from the point of view of the receiver. Imagine a machine selecting single letters from a pool of letters to build a word of three letters: at the beginning the uncertainty is large for which word will result. After the selection of the first letter the uncertainty is smaller, and when all letters are selected the uncertainty has decreased to zero. Thus, information is additive and consists of the decrease of uncertainty. The uncertainty at the beginning is the larger, the more extensive the alphabet is or the more alternatives exist for the individual positions in a string. To express these relations mathematically, Shannon suggested the following convention to describe the measure of uncertainty for each letter.

For a binary alphabet (consisting of two letters only) the measure of uncertainty for each letter is defined as $log_2 2 = 1$. The unit is called a ***bit***. If there is only one letter in the alphabet there exists no uncertainty regarding the possible selection, and for each letter we have $log_2 1 = 0$ *[bit]*. Generally, the expression $log_2(M)$ can be formulated for the uncertainty $\boldsymbol{u}_i$ which is eliminated with each known (chosen) letter $\boldsymbol{i}$. Therein $\boldsymbol{M}$ is the size of the alphabet under the assumption that all letters are equally frequent.

$$u_i = \log_2(M)$$

We can see that this concept, from which the calculation of quantity of information is derived, implies a **definition** of "uncertainty". It is an applicable definition, but not an empirically uncovered law of nature. Shannon defined a unit to measure the coding effort for cases where letters or comparable signs are transmitted. For information coded in a different manner this concept is not applicable.

The probability $\boldsymbol{P}_i$, that a specific letter $\boldsymbol{i}$ reaches the receiver by chance, dependent on its relative abundance, is $\boldsymbol{P_i = 1/M}$, if $\boldsymbol{i}$ is represented in the alphabet $\boldsymbol{M}$ only once. This equation describes the situation in Fig. 87, where one letter is selected at random from a pool of letters. There obviously exists a correlation between the *information*

content of a letter *i* and the probability P_i. For u_i we get the following expression:

$$u_i = \log_2(M) = -\log_2(1/M) = -\log_2(P_i)$$

Thus, u_i describes the "moment of surprise" or the decrease of uncertainty, which is caused when the receiver gets a specific letter. When *P* is very small for a specific letter, the surprise is very large when the letter appears. Shannon's ***H***-function is the weighted average of the uncertainty of the individual states $u_i = -log_2(P_i)$. The derivation of this expression can be explained as follows (Schneider 1996): when a string is ***N*** letters long and contains the letter *i* of the alphabet *M* with the frequency N_i, there are N_i cases in which the surprise is u_i. The average amount of surprise for a string can be calculated as follows

$$\frac{\sum_{i=1}^{M} N_i u_i}{\sum_{i=1}^{M} N_i} = \sum_{i=1}^{M} \frac{N_i}{N} u_i$$

In case of an unlimited number of letters in the string, the quotient N_i/N becomes P_i, the probability for the letter ***i***. Replacing N_i/N in the above formula with P_i and writing $-log_2(P_i)$ for u_i, the formula becomes

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i \quad \text{(bits per letter)}$$

This formula requires that the probability, P_i, that a certain letter *i* is selected is known. It does not state anything about the probability that a selected letter is transmitted successfully and therefore it is only valid for a transmission free of noise.

The amount of transmitted information *R* is the difference between the expectation H_{prior} before transmission and H_{after} after; it is equal to the decrease of uncertainty.

When only the transmitted letters or bits are counted towards the measure of information, the term information refers only to the system "sender of letters" – "receiver of letters". The system-specific reaction of the receiver would be the reconstruction of the letters (not the interpretation of the meaning), which can easily be quantified and compared to the original used by the coding unit. However, if we consider the system "science journalist transmitting information" and "audience in front of the TV set", the system-specific reaction is not the exact copying of letters but the reconstitution of logic relationships between statements. How well the information was transmitted in this case could only be described with the quantitative comparison of correct links between statements, but not with the amount of letters and words that have been transmitted.

In systematics the notion of information becomes important in connection with the question whether similar structures seen in different organisms are homologies or not. Is the content of information of the characters sufficient to assume a homology (implying the existence of an ancestral prototype)? In the same way the question concerning the monophyly of a taxon arises. The credibility of a hypothesis of monophyly depends on the information content or the "value" of the characters which have been used as evidence. Here information is the "trace produced by phylogeny", the system-specific reaction is the identification of homologies and of monophyla. The way in which this problem can be resolved is discussed in chapter 5.1.1.

To quantify the information content of an organ or organism is senseless because it is not known who the receiver is and which process of selection or of "sending" of information is referred to.

1.3.7 What is a character?

Even when observing a single organism we often summarize its properties and classify them: the terms for the animal's colouration ("fur with zebra-like stripes"), the length ratio of front and hind legs ("giraffe-like legs") and so on are already abstractions. Each of these notions represents a group of properties or "characters". Before we use such abstractions each character analysis has to start very critically with the analysis of the material properties of individual organisms.

For the systematist, the "characters" of organisms are of special significance, they are the decisive data used to infer phylogenetic relationships. Basically everything we perceive is a character: a physical structure, movements (behaviour) and songs, molecular components detected with bio-

chemical methods, etc. But caution: a character can also be something we constructed in our heads as a result of comparisons, like "the seven neck vertebrae of mammals", which may have a varying size and shape in different individual organisms. Talking about "characters" it has to be clear which notion we are referring to, because the ontological status can be very different:

a) an observable property of a single organism,
b) a congruent feature we notice in several organisms (a similarity),
c) a homology.

Character (a) refers to a real property or structure the object possesses independently of any observer. The property can, but does not have to, occur in other organisms. It can, but does not have to, be inheritable. The systematist usually works with characters of type (b) and (c). Characters (b) and (c) are constructs (abstractions) of the observing subject: a **congruence** is neither a thing nor a process taking place outside of our consciousness, but the result of a comparison performed in our brains, even if measuring devices are used to do it. A described congruence can also be based on an error, for example, an inaccurate observation or measurement, it is a "datum", not a "factum" (Mahner & Bunge 1997). The notion "**similarity**" is especially problematic: how much congruence has to be present to name objects "similar" depends on the respective experiences and objectives of the observer. A **homology** (character (c)) is a complex hypothesis (ch. 4.2) that refers to perceived similarities and presupposes the inheritance of the same "genetic information". In contrast to the structure of DNA sequences, morphological and physiological characters as well as behaviour are directly influenced by environmental factors. Therefore it has to be examined carefully which properties are inheritable and which not. The song of several bird species is species-specific and obviously inherited. However, in species which imitate others, several song elements are not heritable (as in Old-World-Warblers (*Hippomanias* sp.), or in the European shrike *(Lanais collusion)*).

For this reason it is justified and necessary not to treat characters (= the empirical "data") as given "facts", but rather to test them when doubts arise. A fact can only be a material thing being momentarily in a specific state, or a specific event (Mahner & Bunge 1997). On the contrary, a similarity or homology is a mental object, a construct. Cladists should not make the mistake of viewing character tables as the sacrosanct starting-point of an analysis. It has to be recognized that each statement on the characters of two or more objects is an abstraction. Statements such as "homoplasy in the evolutionary process takes place when …" (Archie 1996) contain an illogical equalization of a construct ("homoplasy") with a material reality ("evolutionary process").

Statements on characters of type (b) (similarities) are intersubjectively testable through measuring or counting. Thereby we gain the necessary confidence that the objects really show congruencies in the described properties. In practice, this inspection often is a methodologically unproblematic though time-consuming process.

Furthermore, it has to be remembered that our use of language is inaccurate: the statement "a group of organisms or a species has a specific character" could be understood in such a way that an object "group" shows specific real properties. However, the "group" as well as the "character" are abstract concepts. In practice it is intuitively understood what is meant (e.g., a homology hypothesis). When a "specific sequence" is mentioned, a specific series of nucleotides found in a single cell could be meant, but also a presumably homologous DNA region showing some variation between organisms. The meaning can be inferred from the context.

Characters in the sense of **properties** only exist for individual organisms, not for groups. A single animal can show new structures in comparison to the parents (e.g., an increased number of sensory hairs) due to a mutation. The mutation does not affect the whole population. However, it happens that a novelty spreads within a reproductive community, or in a clonal population, in such a way that after a few generations each individual shows this novelty (see spread of resistance in bacteria, parasites, and plant pests). Only when this character has substituted completely the previous state in all individuals of an isolated population, or of a species, it is of special interest to the systematist, because it then will be inherited by all descendants, and thereby this **substitution** became an "**apomorphy***" (see. ch. 4.2.2) or **evolutionary novelty**, allowing the scientist to identify each individual as a member of this group

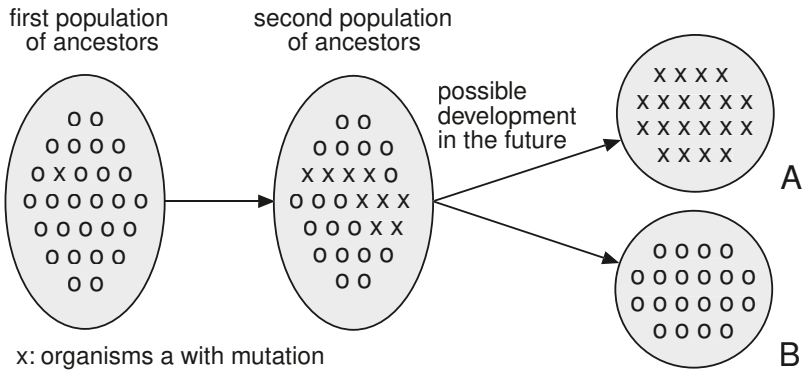


Fig. 5. Alternatives for the dispersal of a new character (a mutation, a new allele) in a chronological series of populations: complete substitution (case A) and loss of the novelty (case B). Populations with polymorphisms (second ancestral population in the figure) represent an intermediate snapshot in a process leading in the long run to the fixation or to the loss of the novelty.

(Fig. 5A). The older state, which has been replaced by the novelty is the "**plesiomorphy****". From the point of view of a scientist, the novelty is a "**supporting character**", because discovery of it allows the determination of group affiliation; it supports a hypothesis of phylogenetic relationships (further explanations see ch. 4).

Polymorphisms in populations are problematic when used as characters for a phylogenetic analysis, unless they are taxon-specific and linked to caste or sex (valid for morphological characters). Character states used to characterize terminal taxa should be unequivocally identified as evolutionary novelties. The simultaneous existence of plesiomorphic (older) and apomorphic (new) states of a single character in one group of organisms does not allow the substantiation of a hypothesis of monophyly on the basis of this character.

Polymorphisms can also be hidden if they were present during the speciation event and were partially lost afterwards (Fig. 6). For phylogenetics the decisive question is whether the divergence of two or more variations of a character occurred before or after the splitting of the corresponding ancestral species into daughter species. If the "splitting of character states" occurred previous to the speciation, and both daughter species inherited the polymorphism, symplesiomorphies can support wrong monophyla (Fig. 6). In this context it is important for a molecular systematist to differentiate, in the case of gene duplications, between orthologous and paralogous genes, as only the analyses of orthologous genes are suitable for the reconstruction of phylogeny (see legend of Fig. 7).

Following the divergence of characters of recent species or of isolated populations backwards in time, one will encounter that population in which the character divergence started. For this time span the term "**coalescence time**" is used (period of time following time backwards until diverging lineages merge; see Figs. 6, 7).

The estimation of divergence time is of interest for the analysis of the genealogy of DNA sequences (see ch. 8.2). The reconstruction backwards in time is based on assumptions about the course of sequence divergence within populations, as described with the "**coalescence theory**", which will not be treated in more detail here. The spreading of alleles (Fig. 5) has to be analysed with the methods of population genetics: a new variant of a character can persist in a population alongside the old one, but it could also be

* *lat.* **ap**ex: point, high cap, crown; *greek* **morph**osis: formation, configuration.
** *greek* **plesio**s: close, neigbouring or neighbour, comrade.

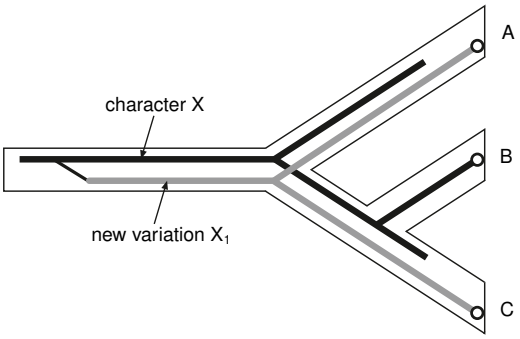


Fig. 6. Inheritance of polymorphisms: if two states (variants "black" and "grey") of character X are inherited by the descendant species, they frequently are not suitable for statements about parentage. Species A and C are not closely related due to the shared novelty X_1 which is lacking in B. X_1 is a shared old character state (a plesiomorphy). The phylogenetic tree of the genes and the one of the species differ.

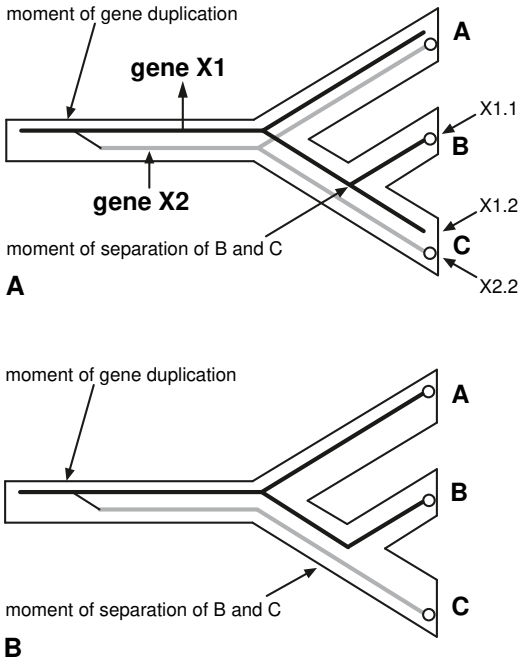


Fig. 7. A: The estimation of the divergence time of species B and C can be erroneous when genes that do not belong to the same "lineage" are compared (genes of the same lineage = **orthologous** genes; see ch. 4.2.1). Orthologous genes are **X1.1** and **X1.2**, a **paralogous** pair of genes however, consists of **X2.2** and **X1.1** in this example. The divergence time of the paralogous genes of B and C is much longer than the time since the splitting of the species which is equal to the divergence time of the orthologous genes. **B:** Differences between gene phylogeny and species phylogeny can also be caused by **lineage sorting**. The outer lines of the tree indicate the population history, the inner lines represent the phylogeny of genes.

that in physically separated populations different character states dominate. The development of allele frequencies through time depends on selection pressure, genetic drift, migration between populations and recombination. Furthermore the rate of divergence depends on the mutation rate. Mathematical theories to reconstruct points of divergences were, among others, developed by Kingman (1982), Kaplan et al. (1991), Takahata & Nei (1990), Hudson (1993), Fu & Li (1993) (see also Li 1997). (Note: the popular term "**coalescent process**" for the fusion of evolutionary lineages in the direction against time is an unfortunate choice, because in our world there exist no real processes running into the past.)

In the practice of phylogenetics, polymorphic characters are rarely a source of errors. There are two reasons for that: 1) experienced systematists choose genes or structures for the analysis which evolve slowly and thus have little intraspecific variation and retain evolutionary novelties for a long time; 2) each polymorphism in the sense of "appearance of gene variants in a population" is only a temporary state existing as long as alleles are lost or fixed. The latter case is a substitution (Fig. 5). Considered over the long periods of time in which systematists are interested, most mutated alleles are lost in the succession of generations.

Morphological structures, such as cell organelles, organs, or external parts of organisms are usually complex, in the sense that they are composed of substructures. It can be assumed that the developmental construction of the visible components of an organism is influenced by a number of different genes. A large number of different mutations can occur which are "neutral", and thus have no effect on the morphology or function of the structures, and are not subject to selection (see theory of neutral evolution, ch. 2.7.2.2), whereas other mutations have consequences influencing the adaptive value of structures. Thereby, a morphological structure can vary in a number of ways. In practice, the complex structure itself (e.g., the

mammalian dentition) as well as a new detail of it (e.g., a tusk) are both called "characters", leading to linguistic inaccuracies and misunderstandings. This problem is treated more extensively chapter 4.2.2.

Sometimes **discrete** and **continuous** characters are differentiated. This can mean that characters can either occur in finite alternative states or in a range of states. To discern between qualitative and quantitative characters is of no use in the discussion about the quality of characters for phylogenetic analyses. At closer view there is no clear difference between "qualitative" and "quantitative". A ratio like "femur twice as long as" or "5 instead of 2 antennal segments", like any other character, has to be evaluated according to whether the character can be homologous or not. There are characters which on principle cannot be conceived as discrete units, such as immunological distances (ch. 5.2.2.5). In this case differences are quantified which become noticeable only as a whole, an effect of the sum of details, but the single evolutionary novelty cannot be identified or homologized. Here the question is whether a signal is present that is probably not the product of chance.

Classes of characters a systematist has to distinguish and further explanations concerning the corresponding terms are presented in ch. 4.2.

Character: perceived property of a material object. Or: perceived identity of two or more objects (construct, result of a comparison).

Substitution: mutation that spread and then dominated in a population and finally replaced or supplemented a previous state.

Apomorphy: a new modification of a character or a new character ("evolutionary novelty"), a substitution or a series of substitutions, occurring for the first time in a specific ancestor population and in corresponding descendant populations. An apomorphy is always named in relation to a group of organisms.

Plesiomorphy: character in an older state (or absence of a character) previous to its modification (which is the origin of an evolutionary novelty) or previous to its replacement by a novelty. This state can only be named in relation to a group of organisms.

Polymorphism: simultaneous occurrence of different variants of a character within a population or species.

Signal: traces which have been left in the hereditary molecules of organisms by phylogenetic processes are called "signal" in phylogenetic systematics. Evidence for the existence of these traces can also be found in the morphology of the organisms. (For the distinction between homology signal and phylogenetic signal see ch. 4.2.3).

1.4 Acquisition of knowledge in sciences

1.4.1 What is a "truth"?

Natural scientists strive to find "the truth" and to make it known, and therefore, they propose statements about circumstances, objects or processes they have observed. The notion "circumstance" (Seiffert 1991: "how things stand") does not necessarily imply a statement concerning nature; the "circumstance" could also be an error or a train of thought. A circumstance that is not existing is, for example, described with the statement "all trees have needles". As this circumstance does not exist (assuming we can rely on our senses), the statement is not true. Therefore it has to be the aim of natural sciences to test which circumstances exist and are "factums" or "facts" in the strictest sense, in order to be able to make statements about nature. Thus a "truth" would be a statement that we are convinced describes an existing fact (**notion of truth in correspondence theory**). The notion of "truth" in logic does not require this correspondence between statement and fact: here it is only significant that statements are linked according to the laws of logic (**notion of truth in coherence theory** = validity of an argument) independent of whether the statements refer to an existing fact or not. The same holds for mathematically founded conclusions. It becomes apparent that nature cannot be explored with the truths of logic or mathematics alone.

These considerations imply that what we call "truth" is not the reality itself, but the statement which we believe describes aspects of reality with

the greatest probability. Therefore it can be easily inferred that each "truth" of science is a hypothesis. Why we can be sure that we have inborn abilities to identify correctly some facts is explained by evolutionary epistemology (ch. 1.5).

1.4.2 Deduction and induction

Forming of hypotheses

Next we have to ask how natural sciences identify what is said to be a fact. This can only be answered for each discipline of science separately, because each area of research has its own methods. Generally an "**existing fact**" (= fact, factum) is the object of a statement formed through scientific methods. The systematists have to be aware that they arrive at hypotheses inductively starting with single observations. This is why absolute truths cannot be reached despite of the use of logic and mathematics (see also chapter 1.4.7).

Natural scientists strive for the description of perceived facts in such a way that a competent reader "understands" what has been perceived. Complicated connections are depicted economically in abbreviated forms, for example, with formulas. The corresponding statements are understandable if the terminology, symbols and other conventions are known, and it should be clear with which logical connections one statement was deduced from other ones. The inference of a statement from more general and simpler statements is a **deduction**. The backward reconstruction of a deduction that shows how a complicated statement can be understood or how it can be derived from simpler ones is called **proof** in mathematics (regressive deduction). Looking for the roots of all proofs one encounters the sensory perception and axioms or first principles. These **axioms or first principles** can be the **unsubstantiated starting statements** of an otherwise strictly logical series of statements. Here unsubstantiated only means that these principles *cannot be proved* within the limits of deduction. Starting statements (axioms) can express conjectures (as for example, assumptions on the existence of the "molecular clock", a prerequisite for some mathematical methods used to estimate genetic distances: ch. 8.2.2), or relate to experiences of everyday life (e.g., the visible and measurable phenotypic variability within reproductive communities as basis for statements on selective advantages of character states). This relation to experiences is analogous to the rooting of scientific terminology in colloquial language (see ch. 1.2). It is easy to comprehend that logically impeccable deductive reasoning is of no use if it is based on senseless axioms or on assumptions not congruent with existing facts. In natural sciences starting statements should be constructed or deduced from elementary everyday experiences. Axioms in this sense are not working hypotheses which can be tested in the framework of a deduction! The extrasubjective reality cannot be reconstructed by deduction alone.

Even though the mathematical methods of phylogenetics (see ch. 6, ch. 8) can be deductively concluded from starting statements, the uncertainty of the truth of assumptions contained in the starting statements remains. It will be shown that such assumptions are important hypotheses (temporary conjectures) in phylogenetic systematics that influence the results.

Hypotheses are statements with which we attempt to explain observations or to deduce general laws from singular observations. Statements contained in hypotheses are never facts (it is however a fact that hypotheses can be formulated). Hypotheses originate in **inductive research**. Existing facts are tentatively identified from single observations in nature or in experiments. One has to ask how probable it is that an assumed fact (process, event) can cause the observed occurrences or structures. A similarity observed comparing two individuals could be the result of inheritance of the same DNA nucleotide sequence. The observation is the presence of a similarity (e.g., the same colour and length of the fur), the hypothesis is the existence of a common ancestor with the same characters. The hypothesis can be "verified" through the collection of more single observations. It could be tested, whether other characters are also congruent, as would be expected in the case of inheritance. To reduce the influence of subjective distortions or biases of perception, other persons can be asked to repeat the observation: the observation is **intersubjectively testable**. The parts of the experience which vary intersubjectively can be deleted from the

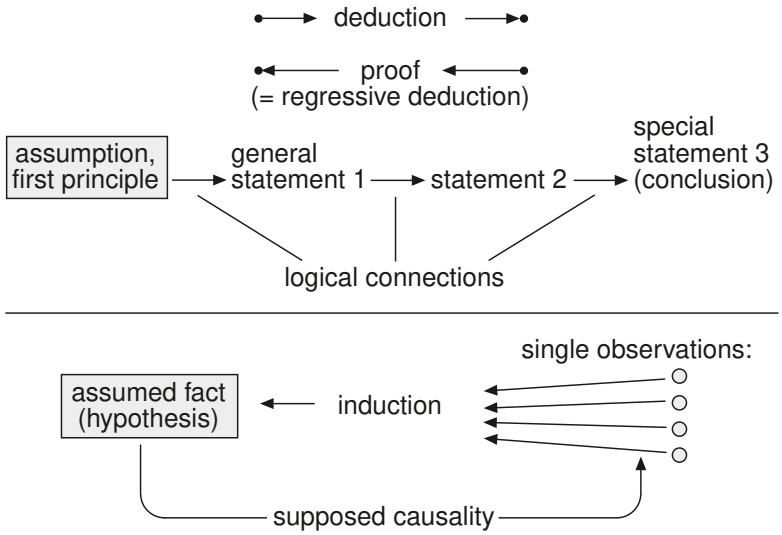


Fig. 8. Deduction and induction. After a successful test the hypothesis has a better support, but it is not proved: in inductive research a proof (regressive deduction) is not possible, but rather a probability statement. It is a question of problems with "inference". (The expression "to infer a phylogeny" indicates more clearly that an inductive conclusion was gained than the wording "to reconstruct a phylogeny".)

protocol of the observations. The "fact" known in this way, however, remains a product of human perception, it depends on experience and measuring instruments. The history of natural sciences teaches that the expected objectivity of results of an analysis often turns out to be an illusion. Knowledge is always hypothetical.

An essential difference between induction and deduction is that in an inductive process the scientist has to select those observations that will be used for the inductive solution of a question. On the contrary, deductive conclusions can be automated because the laws of logic are invariable even though the result depends on the starting conditions (see 1.4.4.): if the premises are true, the conclusion is inevitably true, if deduced logically. Statements on probabilities are out of place here. In induction on the other hand, the premises can be true, and the conclusions wrong. The conclusion "all dogs have brown fur" deduced from 20 observations of dogs that all were brown is wrong. A further difference is that deductively more complex or more special statements can be obtained from simple statements, but inductively the general is inferred from particular instances (single or special cases, samples).

Biology is, as well as physics or chemistry, an **experimental science** that obtains hypotheses using inductive methods. The fact that often strictly logical and mathematical methods are used to process data must not lead us to forget that, especially in phylogenetics, individual observations are premises, and that the applicability of mathematical methods depends on certain conditions (even though often this is not explicitly stated). The application of a mathematical method represents a deductive step, inevitably always leading to the same result when the same individual observation is given. The calculation, however, takes place in the framework of inductive research: the premises necessary for the use of a specific mathematical method as well as the results we obtain are not inevitably truths. Users of cladistics and of diverse methods of molecular systematics must not forget this. For example, the samples used in molecular systematics are sequences used for the comparison of selected species. The general conclusion then may be that the calculated values are "genetic distances" (see ch. 8.2) representative of a group of species and a historical period of time. As soon as statements on phylogeny were deduced from these values in the next step, the transition to hypothetico-deductive methods occurred (ch. 1.4.3).

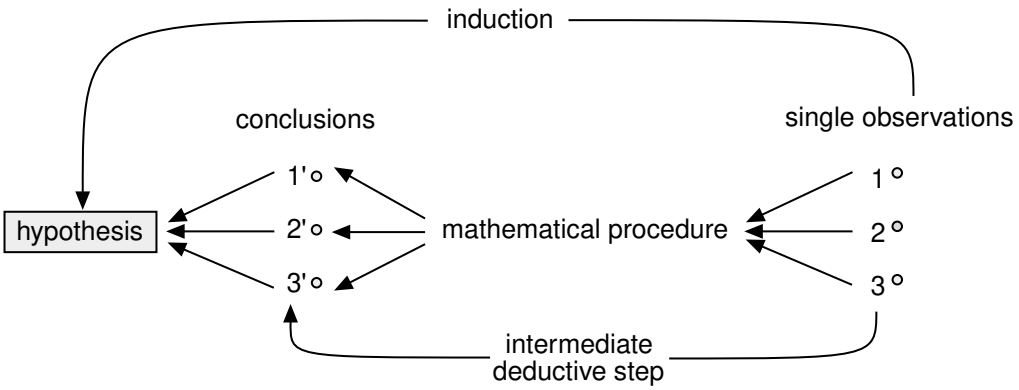


Fig. 9. Deduction within an induction. In the example of the calculation of genetic distances, it is important to comprehend that the premises for the deductive step, namely the hope that the selected species and characters represent a suitable sample, are at the same time premises for the whole induction. Also, the basic conditions for the deduction, for example, the assumptions implicitly contained within an algorithm, are at the same time conditions for the quality of the final hypothesis. When we calculate genetic distances using the very simple Jukes-Cantor model of sequence evolution (see ch. 14.1.1), the basic condition or first assumption is that the model represents correctly the historical processes of sequence evolution, the variations of substitution rates. The results of the deductive step (the calculated distance values) are always logically correct ("calculated correctly according to the formula"), and independent of whether the basic conditions correspond to reality or not. The obtained hypothesis does not have to be correct.

Often the **reproducibility of results** is used to "prove" a hypothesis to be probably correct. If several datasets are used independently for calculations (single observations 1-3 and conclusions 1-3 in Fig. 9) and the results agree with each other, this may only mean that the samples are similar, but not necessarily that the deduced hypothesis is "true" due to the reproducibility of the results. It could be, for example, that three different character sets are informative, but the selected species do not represent a suitable sample (see "symplesiomorphy trap", ch. 6.3.3). Also, the repetition of results (e.g., reconstruction of branch lengths using the Jukes-Cantor model) is not a test for the correctness of the basic assumptions (implied with the Jukes-Cantor model) either, because the deductive step principally never tests the correctness of premises and basic conditions. The premises have to be tested with a specially adapted and independent method.

Sober (1988) distinguishes between **induction** and **abduction**. In this case the notion induction is restricted to the generalized statement on properties of a set of objects, derived from a sample (e.g., "all ravens are black"). Induction in this sense requires an assumption on the uniformity of all objects of the set (Hume's "Principle of the Uniformity of Nature" (1777)). The abduction is the inference of an explanation, a hypothesis on mechanisms or the reconstruction of a cause, also starting from samples. According to Sober only the abduction requires the assumption that the most parsimonious explanation is the most probable one. This opinion has to be criticized, because induction in the sense of Sober also implies assumptions on causes or mechanisms: the statement "all ravens are black" obtained from samples is with higher probability right than the statement "all bikes are black", because the colour of ravens can have only one common cause (the descent from black ancestors), whereas bicycles are manufactured in different factories and are painted individually. Thus here also a most parsimonious explanation is wanted. The term induction as used in this book includes the meaning of the notion abduction.

1.4.3 The hypothetico-deductive method

Scientific progress is neither possible through pure observation of nature (empiricism) nor by thinking alone (rationalism). Each scientific theory is based upon hypothetico-deductive steps and consists of several hypotheses. The hypothetico-deductive method requires:

1. a primary hypothesis in form of assumptions (axioms, postulates) mostly obtained inductively, from which
2. a prediction is obtained deductively.
3. The prediction can be tested empirically, independently of the method which led to the

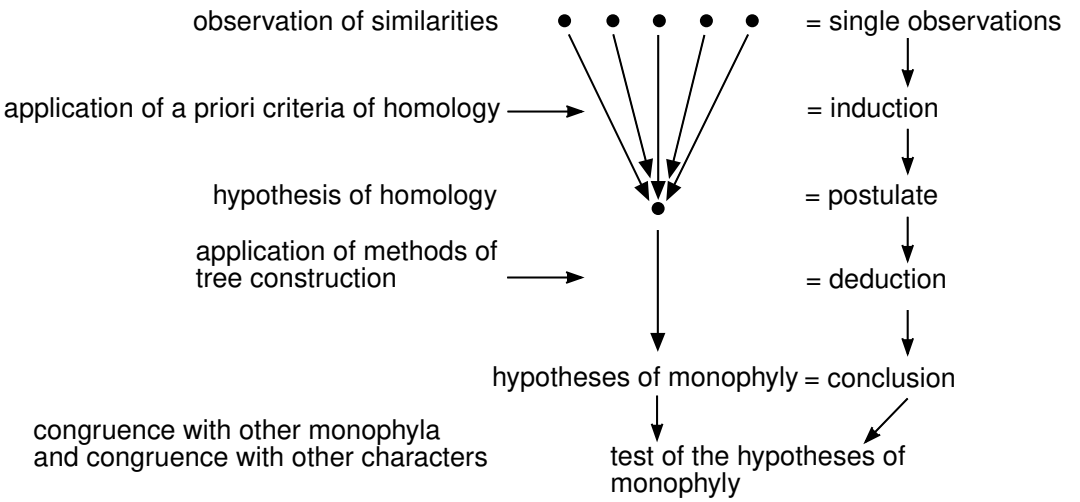


Fig. 10. Hypothetico-deductive reconstruction of phylogeny. In phylogenetic systematics, what is finally found and tested are the hypotheses on relationships. A prediction is that with another suitable set of data the same phylogeny should be reconstructed. As these hypotheses are dependent on several premises (quality of species samples, quality of selected characters, first assumptions implied in the algorithms used; see ch. 6.1.11, ch. 9), a convincing verification is only possible with analyses using different premises (different samples of species, other characters, other algorithms). If in a second analysis even one of the premises used for the prediction remains the same, a congruence with previous predictions could be caused by the same bias or erroneous assumptions. Sometimes it is surprising that the same, clearly wrong topology (e.g., with polyphyletic molluscs), is repeatedly obtained, although each time other genes have been used for the reconstruction. This may have different causes: the same sample of species, the same algorithm has been used, possibly the same genes have been used from different species, but evolving with the same heterogeneity of substitution rates in related species, and which therefore, may not be suitable for the detection of geologically old divergences.

prediction. If it is verified the axioms will be considered as confirmed.

A **prediction** is a hypothesis, which can go beyond the empirical experience and is emendable like any other hypothesis (Bunge 1997). The context shall be illustrated with an example (Fig. 10):

Inductive step: a hypothesis of homology may be proposed after identification of a structural similarity in two organisms that is too complex to be the result of pure chance. In doing so, one has to be aware of the criteria necessary for the estimation of the probability of homology (ch. 5.1.1). Is the homology in the organisms that are being compared unique in nature? The hypothesis follows that the similarity could be a shared evolutionary novelty.

Deductive step: within the bounds of a phylogenetic analysis the statements on homologies serve as postulates, from which it can immediately be deduced or predicted that organisms sharing a unique character assumed to be a homology should be more closely related to each other than to those lacking this character primarily (i.e., in the latter group the absence of the character is not the result of a reduction). This step of deduction follows, among others, directly from the laws of classical genetics. The homology hypothesis is always inevitably interlinked with a hypothesis of relationships. If the character is an evolutionary novelty, an inescapable hypothesis of monophyly follows.

Prediction: further characters not yet studied will be congruent in the species of the monophylum.

Test: a phylogenetic analysis carried out correctly with other informative data serves as a test for the prediction (see above: reproducibility of results, ch. 1.4.2). The test can lead to a weakening or rejection of the hypotheses of homology and monophyly. Other testable predictions are possible, for instance on the geographic distribution or on the geological age of groups of species.

1.4.4 Laws and theories

We have to be aware of the fact that we are always working with hypotheses. When these have been verified several times, and if we found that these are valid not only for a single historical event but for recurring processes, we may call them **laws**. A phylogenetic hypothesis (e.g., "the monophylum Monotremata has to be placed at the base of the phylogenetic tree of recent Mammalia") only applies for the individual case (the example refers to a single specific speciation event in the early history of mammals) and is not a law in the normal sense. Laws allow predictions. It is a law that members of a monophylum (see ch. 4.4) show some characters not occurring outside this monophylum, because these characters (apomorphies) originated in its "stemline" (ch. 3.6). It can be predicted that an unknown organism belongs to this monophylum if it shows these apomorphies. Furthermore it can be predicted that in this case the unknown organism will also show further characters not yet studied which occur in other members of the same monophylum.

By combining several laws to a more general or superordinate statement we formulate a theory. Theories which have often been confirmed or strengthened and where further testing does not seem to be necessary are sometimes called **paradigms**. The generally high regard a layman has for laws and theories of natural sciences must not mislead us to forget that we are dealing only with more or less well supported hypotheses. This uncertainty also exists for statements on causality.

Therefore, in biology, as well as in other empirical sciences, we are not allowed to formulate our statements as absolute "truths", in the form of "it is so" – sentences. Rather it should always be pointed out that some **evidence** supports a hypothesis ("most likely it is so"). As observations in biology and medicine usually show a much higher variance than in physics or chemistry, mainly due to the intricate complexity of the systems analysed in biology, hypotheses are less reliable and often do not allow predictions. If the number of independent observations supporting the same hypothesis is very large, as in the case of evolutionary theory, the probability that the assumed circumstances really exist is also very large.

Hypothesis: a statement on a circumstance reconstructed from observations or experiences (a conjecture).

Law: a hypothesis confirmed several times; its predictions proved to be correct so far.

Theory: superordinate statement that combines several laws.

Paradigm: generally accepted theory which forms the unquestioned starting point for new hypotheses.

Proof: chain of logically interlinked sentences with which a complicated statement can be traced back to simpler, understandable statements.

Evidence: observations supporting a hypothesis.

1.4.5 Probability and the principle of parsimony

Probability statements are only possible for processes

As we often make statements on whether a hypothesis is "probably true" or "probably false", at this point we have to discuss the notion of "probability". Probability statements belong to inductive research, statistics is the set of methods with which these statements can be obtained.

According to the considerations of Karl Popper, a probability statement always refers to real events or processes (Popper 1934, Mahner & Bunge 1997), but not to theories. Events "probably happen", theories, however, are verified. Fig. 11 visualizes this circumstance.

Events can be the result of processes that modify the properties of objects or the composition of groups. These could be natural phenomena, but also the collection of samples, or cognitive processes. Probability statements are only possible if processes run **stochastically** and when they are observable with all relevant parameters. The radioactive decay of an element, for example, is such a process. No statement is possible concerning the precise moment of decay of a specific atom, but good predictions are achievable on the expected average for a large number of atoms, resulting from the statistical half life of the isotope. **Deterministic** processes always end with the same results. Drop a stone: it always falls

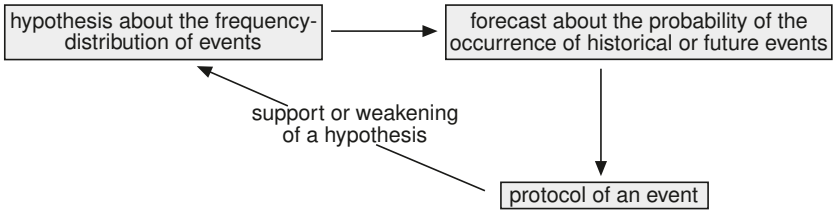


Fig. 11. Popper's conception of the support of hypotheses.

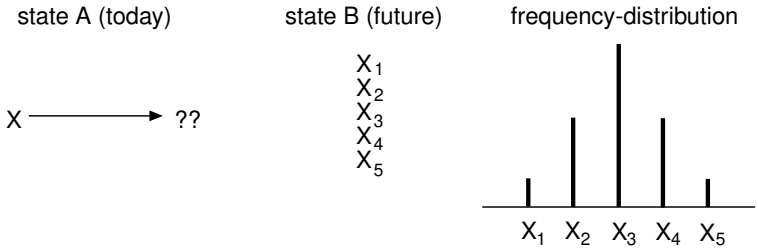


Fig. 12. The **probability of events** is predictable in stochastic processes, when the frequency distribution of possible results (state B) and the starting conditions (state A) are known. In the reverse case, the probability distribution of the starting state can be reconstructed from the resulting states, when the course of the process is known.

towards the center of gravity. The result can be predicted if the laws valid for the course of the process are known. Scientists call some processes chaotic because they develop unpredictably for our eyes even though they can be described using simple mathematical formulas. Such processes are predictable if the starting state and the factors influencing the process are exactly known. In **chaotic** processes the result is not predictable in practice. Chaos in this context means that a process is extremely sensitive to starting and marginal conditions, which cannot be grasped exactly enough to predict or calculate the result. A typical example is the long-term weather forecast (e.g., guess in September the weather for Christmas Eve in Berlin), which is very imprecise in contrast to a forecast of a solar eclipse. Evolution is the visible sum of processes which can be assumed to take a partly stochastic, partly deterministic or chaotic course.

Stochastic evolution: the course of evolution is determined by random events which show a lawful accumulation of frequencies of alternative final states.

Deterministic evolution: a specific result develops inevitably when specific initial conditions are given.

Chaotic evolution: evolution proceeds in an unpredictable way, because the factors determining the occurrence of random events are numerous and/or complex and very variable.

Hypotheses on the course of stochastic processes allow prognoses on the expected changes (Fig. 12), if the starting state and properties of the process are known. When the result of a process is known and there exist substantiated statements on the most probable course of the process, it is possible to estimate the most likely starting states. In this way the fraction of radioactive isotopes in a sample of carbon can be used to calculate the age of old fossil organic matter (^{14}C-method).

Usually, for evolutionary processes only the final states are known (e.g., characters of terminal taxa), the historical process itself has not been observed. Methods to reconstruct phylogenetic trees which rely on models for the process of character evolution require assumptions on the course of the process to find the most likely starting state of character evolution or the most probable course of character evolution (ch. 7, ch. 8). Assumptions on the course of the process can, for example, be statements on the mutation or substitution rates, statements on the rate of character changes

real event: state A → process → state B

possible analyses:

state A	course of process	state B
known	known	predicted
reconstructed	known	known
known	reconstructed	known
reconstructed	estimated	known

Fig. 13. Possible analyses of stochastic processes. The probability that a real event takes place can be estimated when the starting or terminal states and the course of the process are known, or when substantiated assumptions on the course of the process exist. "Known" means that the state or the process was observed in nature or could be reconstructed from evidence or from samples. The quality of implied assumptions determines the reliability of conclusions.

(change of leg length or of ovary size per unit of time), or statements on selection probabilities for alleles. These assumptions can be obtained through comparison of final states, whereby the risk exists that the assumptions are not in accordance with reality.

Probability statements do not explain the causes of events

Probability statements do not allow the derivation of a **substantiation** for singular historical events. Thus for the radioactive decay of an isotope a statistical statement can be made, however, it is not possible to find out why a specific atom has reacted at a given time.

Processes exist outside and inside of a subject

In Fig. 14 scenarios are summarized for which an observer can make statements on the course of a process. In nature there are processes existing outside the thinking humans, but there exist also cognitive processes. Being a subject one can find oneself in three different levels:

Level A: as observer of processes occurring in nature.

Level B: as a subject who analyses the traces left by processes to deduce hypotheses on their origin or on the course of the process, or who analyses starting states and formulates prognoses on possible future events.

Level C: as observer of another subject formulating hypotheses.

Thus, as a subject, I can make different probability statements when historical events, such as the evolution of organisms, are examined:

Case A: the estimation of the probability that a specific process occurred or will occur (probability of events, see Figs. 12, 13) relies on hypotheses about the frequency distribution of events. When for example the process parameters and the final states are exactly known, a probability statement on the most likely starting state is possible (see. ch. 8.3, maximum likelihood methods).

Case B: the estimation of the probability that patterns or clues observed by me are traces of a natural process can be independent of the course of this process: if I identify several complex identities in two organisms, I can deduce a homology statement without having to infer the precise course of evolution. The identification of patterns is a cognition process. As this process takes place within myself, I subjectively estimate the probability that I recognized and evaluated the observed identities (characters) correctly.

Case C: estimation of the probability that a fellow scientist has erred or reconstructed a process correctly. Hereby I consider the "**quality of the receiver**", as for example the education and experience of the scientist and the accuracy of his or her research, and the "**quality of the trace**", thus the number and quality of the clues the scientist has analysed. In this context we also estimate the probability that a hypothesis makes a correct statement about a real fact. Being an observer I follow the cognition process of a third person. I try to evaluate her or his cognition process objectively. To do so, experiments with test subjects are possible, in which the observer knows the real facts that are being analysed by the test person.

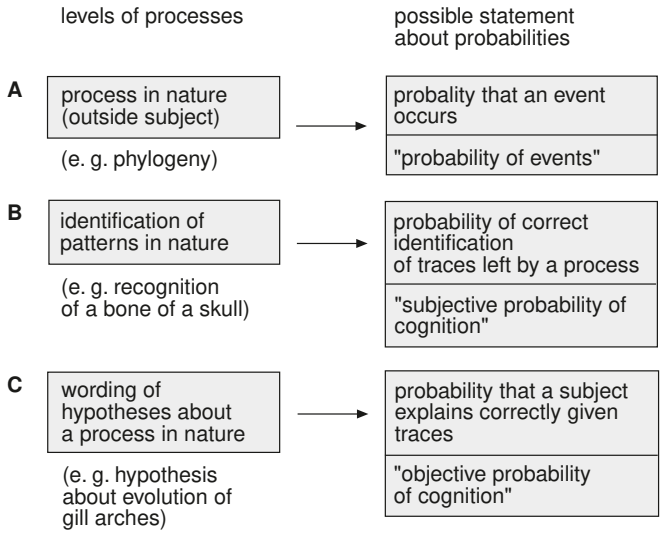


Fig. 14. Scenarios in which processes can be observed and probabilities can be estimated.

In practice, a scientist should consider all three levels. Thus we should ask whether we have the relevant education and experience, and thus place ourselves in level C. Furthermore, we have to evaluate the quality of the clues or of possible traces (level B) and at the same time we should ask whether there may exist in nature processes creating such traces (level A).

Probability statements on hypotheses

It cannot be denied that in a group of alternative hypotheses interpreting the same circumstances we can rank them, differentiating those that are "probably true" and others which are less in agreement with the known observations. So, of the following two hypotheses only one is "with greater probability true":

a) "the lens eyes of cuttlefish are homologous to the lens eyes of vertebrates".

b) "the lens eyes of cuttlefish are convergences to those of vertebrates and are similar only by chance or due to the action of similar shaping forces".

The question arises to which process such a probability statement refers.

The statement "hypothesis b) is more probable" does not mean that the historical event "lens eyes evolve by chance to a similar shape" is more probable, which would be an evaluation of the probability of events, but it implies that a lot of information in favour of hypothesis b) should be present (evaluation of the probability of cognition). In practice we do not estimate how probable it is that a retina can evolve from epidermal or from neuronal tissue in a given time interval (this is currently not possible for such a complex evolutionary process), rather we evaluate the complexity of the visible patterns. A theoretical foundation for the differentiation of the quality of homology statements is presented in ch. 5.1.

From the point of view of the acting subject, hypothesis b) is the better supported one. From the point of view of an observing, all-knowing third subject placing at disposal clues to the scientist, the true fact will be detected with higher probability if the scientist identifies many details of the clues, in contrast to one that spends little time for the examination of available tracks. If we make a probability statement in the sense outlined above, we do as if we were the observing third (level C of the preceding paragraph), although we do not know the true fact. We can assume that out of 100 equally and adequately trained scientists a high portion will always reconstruct a historical event correctly if a sufficient number of traces of the event are present.

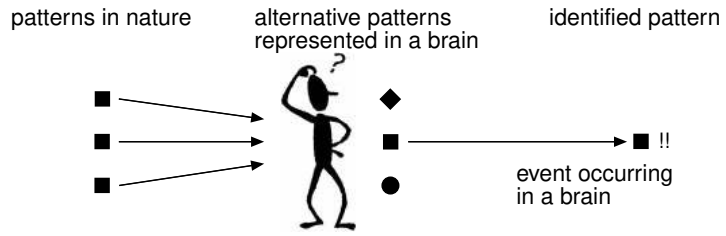


Fig. 15. Gain of cognition as a special case of an event.

Classes of probability statements

For the methodology of systematics we have to differentiate two different **classes of probability statements**:

a) estimations of the probability that an event takes place in nature or that a certain process develops. These are statements on the **probability of events** (natural probability, also misleadingly called statistic probability). It depends on the conditions existing in nature and is independent from the observer. So the probability that inheritable diseases appear *de novo* in a person can depend on the frequency with which mutations hit a specific gene.

b) estimations of the probability that a specific process will be identified correctly from a given observation. This **probability of cognition** (Fig. 15) presumes the existence of a subject and depends on the **quality of the available data** and on the **data processing**. "Data quality" means that the sample has to be representative and the individual data should be informative, containing no or little noise. Without knowing the process of evolution, every systematist has to evaluate whether similarities could be homologies or not. "Quality of data processing" means that, for example, the subject should have adequate training. (Perception, cognition, and the scientific gain of knowledge are also natural processes which can be evaluated objectively. The differentiation between probability of events and probability of cognition serves only methodological intentions.)

The situation illustrated in Fig. 15 presumes that material things that are similar exist outside the subject (e.g., squares). In the brain of the subject there exist correlates representing different shapes (e.g., diamond, square, circle). The **process of cognition** not visible from outside is the comparison of the patterns coming in through the sense organs with the correlates, the **event** is the identification of a specific correlate. The probability that the identification is correct depends in this example on whether

a) a suitable correlate is present in the central nervous system, and whether
b) the information coming in from the sense organs is sufficiently detailed and representative.

In practice, this context can be seen in the fact that trained specialists (e.g., radiologists) can identify in a picture an object or circumstance (e.g., damaged joints) which a layperson would not recognize. Accordingly, the comparative morphologist in practice can only identify correctly homologies with a high probability when a) the scientist is well trained, b) when she or he has analysed the objects carefully in detail, and c) when the objects have enough visible structural details in common to allow the identification of shared patterns.

Historical research mainly relies on the evaluation of the probability of cognition

For a historical event, the course of the corresponding process can only be reconstructed with precision in clear cases. The process should develop according to known laws, and well recordable final states (results of the process) should be present. The relative position of a planet cannot only be predicted precisely but can also be inferred for the past, though the result will only be correct if all factors are understood. The influence of the gravity of an unknown comet could be a source of errors.

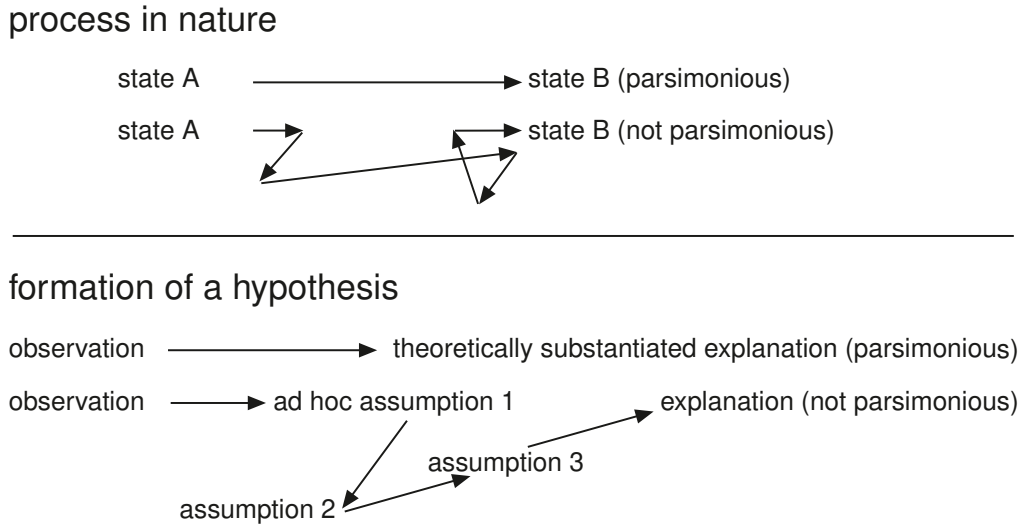


Fig. 16. Parsimony in nature and parsimony of hypothesis-forming. The most parsimonious explanation for the occurrence of **events** is the one that assumes the process course most frequently found in nature. This approach can be used when parameters of the process of character evolution are inferred, a prerequisite for distance and maximum likelihood methods (see. ch. 8).

In historical research, especially the probability of cognition has to be estimated, but rarely the probability of events. A statement on the probability that the impact of a meteorite is the cause of the mass extinction at the Cretaceous/Tertiary boundary does not depend on the probability that such an event really occurs (e.g., estimating how often large meteorites hit the earth). This is an **important aspect of historical research**: even though the theoretically calculable probability that an impact really occurred might be extremely small, possibly close to zero, it can nevertheless have taken place. The probability that we identify the event correctly does not depend on the probability of the event, but on the certainty with which a causal correlation can be concluded from fossils and from traces of the impact. The clearer a connection can be established (e.g., through exact dating of rocks with high iridium content at different places of our planet), the higher is the probability that of many alternative hypotheses a specific one describes the historical event correctly. A further example: a historian would not want to calculate how probable it is that Caesar met Cleopatra and fell in love with her on the basis of a) data on the health state of Julius Caesar, b) the range and security of available transportation vehicles, c) the frequency of meetings between Caesar and females, etc. The historian is much more interested in any evidence that this event really took place and she or he will evaluate the available historical documents.

It is like the evaluation of a distinct pattern: the question whether a pattern (the shape of traces) could have originated by chance as a result of several independent processes or through only one specific singular process is a probability decision (see "probability of homology": ch. 5.1.1). Statements on the probability of cognition make assumptions on the **information content of the available data**.

Probability statements in connection with calculated or **reconstructed trees** can refer either to the probability that

a) the data (characters) used are informative (ch. 5.1), and the reconstruction therefore portrays the real events, or
b) to the probability that specific evolutionary processes took place (see ch. 7, ch. 8).

Most probability statements of cladistics (see, e.g., "bootstrapping", ch. 6.1.9.2) are only useful artifacts of mathematical methods which estimate neither the quality of the data nor the course of the evolutionary processes (see also "deduction" in ch. 1.4.2), but rather the extent of congruence

between data and a topology calculated from them.

The principle of the most parsimonious explanation

The "**principle of parsimony**" (principle of the most parsimonious explanation, principle of the economy of thinking) is a rule, a methodological resource, used for the comparison of explanations (hypotheses). It has proved useful to avoid unnecessary ad-hoc-assumptions for the explanation of observed circumstances. By asking which minimal assumptions are sufficient to explain a phenomenon fanciful stories can be avoided. This rule, "Ockham's razor", is attributed to the theologist and philosopher Wilhelm von Ockham (ca. 1280–1349) ("pluralitas non est ponenda sine necessitate"). The principle does not affect level A) (extrasubjective process in nature) but level C) (formulation of hypotheses, see Fig. 14).

The principle of the most parsimonious explanation does not mean that the preference for the **simplest explanation** is equivalent to the assumption that the existence of a simpler process is more probable than the existence of a more complex one. Evolution is not "parsimonious", it did not proceed according to a given plan, but in a chaotic way. This is why, for example, the human spine has not been designed "from the start" for standing and sitting. This is noticeable in repeated damage to intervertebral discs. The heart of mammals with its twisted vessels is not an optimal technical construction, but a useful solution built from the given starting material. Hypotheses on the course of evolution should not search for the straightest of all possible solutions (e.g., proposing a line directly from fish to whale), but have to reconstruct the often complicated historical events.

The most parsimonious explanations for the **interpretation of identities** of characters of organisms rely on the estimation of the probability that identity in many details can be attributed to a **common cause**: it is more probable that a complex chain of events (the evolution of a complex organ) occurred only once, than that exactly the same sequence of mutations and selection processes occurred several times (see criterion of complexity for homology hypotheses in ch. 5.1).

The most parsimonious explanation for the occurrence of specific identities in a limited number of organisms is that the identities originated in a **common stemline** and therefore are lacking in other organisms. The parsimony method of cladistics used to reconstruct phylogenetic trees can be deduced from this reasoning.

In systematics the principle of parsimony is required in particular for the analysis of characters and for the analysis of congruence of hypotheses of homology.

1.4.6 Phenomenology

In the following chapters the term "phenomenological method" is used, originating from the terminology of inductive research. It means a scientific method, which tries to avoid the use of axioms or first assumptions in order to study first of all what can be seen or experienced in nature or society. Especially those assumptions already implying an explanation of the observed have to be avoided. However, no method is absolutely free of assumptions, the theories of physics, for example, frequently have to be accepted as the basis.

A "phenomenon" is the **perception** of a thing or of a process by a subject (see Mahner & Bunge 1997), such as the sensation of heat at a fire or the observation of beating wings of bats and birds. This perception cannot be equalized with the extrasubjective reality that causes it. Therefore a scientist has to test which real properties or processes correspond to the perception. Perceptions can be tested intersubjectively. In case of subjective sensations (e.g., colour vision) they are comprehensible and the stimuli causing the perception are measurable. The measurement is a different method of perception, often more precise, and more easy to compare intersubjectively

So, the phenomena should be described first. These observations do not have the function of samples: if I observe several mosquitoes sucking blood, I may deduce the statement "the stinging mouthparts of mosquitoes *can* be used to suck blood". With this sentence I do not go beyond the observation, no assumptions are made besides that I want to trust my eyes. For this purpose **statistical calculations are not necessary**. How-

ever, no universally valid law can be deduced from this conclusion. Users of the phenomenological method do not claim to be able to predict "*all* mosquitoes can suck blood". This would be an unsubstantiated statement and it is in fact wrong. Phenomenology is suitable to describe historical events or their consequences, which are singular facts "not obeying any laws". Phylogenetics deals with such historical events (e.g., the colonization of the Galápagos Islands and subsequent speciation events).

1.4.7 The role of logic

Logic places tools at the disposal of scientists and helps to formalize arguments such as the interlinking of statements. Logic requires abstractions from empirical experiences and offers rules which can be used for the correct deduction of conclusions. Logic, however, does not make any statement on the ontology of the conclusions obtained. Despite its abstract rules, it is not a product of fantasy without relation to reality, but a collection of abstractions resulting from the regularly observed relationships of objects of the real world. The everyday experience is the motive to set up rules for gaining valid conclusions:

(a) When it is raining, the landscape gets wet. (b) The landscape is dry, (c) therefore it did not rain.

These rules also exist outside of our consciousness as proved by machines for electronic data processing. An example for the rules of logic is that attributes of a subset do not necessarily have to be present in other elements of the superordinated* set. The argument

"(a) all dogs have fur; (b) dogs are animals, (c) therefore all animals have fur"

is logically wrong. On the other hand the argument

"(a) all insects have wings, (b) fleas are insects, (c) therefore, fleas have wings"

is logically correct although the conclusion (c) is factually wrong!

Starting from premises (presumed assumptions: "all dogs have fur" or "all insects have wings") one arrives at conclusions ("all animals have fur" or "fleas have wings") using the rules for the interlinking of statements. However, the observance of the rules of logic does not guarantee that the logically necessary conclusion is correct, because logic does not make any statement on the correctness of the premises (assumptions, axioms). In the example above the assumption "all insects have wings" is wrong, but the logical interlinking of the sentences correct. The conclusion is not correct because the premises are wrong. The conclusion could have been accidentally right: if the premises are incorrect there is no reliable conclusion. [In the first example ("all animals have fur"), the starting sentence is correct, but the interlinking of sentences does not follow the rules of logic, because it is not allowed to derive from properties of a subset (dogs) properties of the superordinated set (animals). In the second example, the starting sentence ("all insects have wings") is not correct.]

Therefore, applying logic does not guarantee finding the most likely conclusion, an observation which is of relevance to the application of strictly logically constructed methods of research: if the calculation is done with worthless data, the result is also of no value. If the method implies unrealistic assumptions, the result is not trustworthy. The application of logic to gain scientific statements is a **necessary condition** (*conditio sine qua non*), but not a **sufficient condition**. (A condition A is **sufficient**, if the content of truth of A is adequate to prove the truth of statement B that is derived logically from A.)

1.4.8 Algorithms and gaining knowledge

The results of automated calculations are considered to be reliable, because the machine usually works flawlessly with the given algorithms (and only with those). This permits the analysis of large datasets, the testing of a large number of possible solutions, and the choice of those that best match the given data. Nevertheless, it is obvious, that a calculation is not necessarily reliable only *because it was done by a computer*.

* *superordinated*: more inclusive, or higher ranking set.

A scientific examination of the reliability of results has to consider five levels:

1. Checking the reliability of the machine: does it calculate correctly?
2. Checking the programming of the algorithms (detection of programming bugs).
3. Checking the logic of the algorithms: are they suitable to answer the question?
4. Checking the assumptions (axioms) implied by the algorithms.
5. Checking whether the data are suitable to answer the question at hand.

Whereas steps 1 and 2 are more concerned with technical problems and can be tested partially with the computer itself, especially steps 4 and 5 can momentarily only be carried out by experts of the field that master the methods and the actual state of knowledge of their discipline. Often in the beginning it is not known which assumptions an algorithm uses. Furthermore, the best algorithm calculates nonsense if the data are not representative for the case to be analysed (example: plesiomorphy trap, ch. 6.3.3). The often heard proposition, that a computer analysis principally enables a better approach to the truth, is principally wrong as long as the automated machine does not master the methodology and state of knowledge of the corresponding branch of science.

Presently available computer programs can be used within the bounds of inductive research for deductive intermediate steps (Fig. 9), but not to work out hypothetico-deductively a hypothesis. A trained scientist is needed to choose the best set of samples, as well as for the formulation of working hypotheses, and for the examination of the plausibility of results.

As long as it is not clear how correct sampling can be done by robots, how the quality of datasets can be evaluated in an automated way, and under which circumstances axioms can be accepted, it is risky to take a computer program and a machine as a "black box" that reconstructs "the phylogeny" from some input data. For the time being every systematist has to become familiar with the theory of systematics and when using computer programs she or he has to know which principles the "black box" uses and which assumptions are implied with a method.

1.5 Evolutionary epistemology

As already noted (ch. 1.4.2), scientific proofs in the strict sense are regressive deductions tracing back logical conclusions to first sentences. These starting statements represent **unprovable axioms** or first assumptions. In natural sciences to avoid senseless assumptions, axioms are thought to be based on everyday experiences, on perceptions and on the correlation between notions and perception formed during learning of a language. This is point zero of every reasoning. This point zero exists also in inductive research.

A hypothesis may be intersubjectively testable. Hereby we rely on the assumption that other subjects are able to gather experience with their sense organs and brains in an objective way and we compare their statements with our own experience. As it is in principle imaginable that our perception of our environment is only a product of our brain whose abilities includes the comprehension of mathematics and logic, but also the construction of new ideas, material objects equivalent to our conception of the world do not necessarily have to exist. Then, however, what we call "scientific knowledge" is also questionable.

Referring to evolutionary epistemology we get the arguments that allow us to accept the point zero which is the starting point of every formulation of cognition. This concept states that prescientific knowledge portrays aspects of the real world because the **ability of cognition,** which depends on the structure and physiology of sense organs, data processing in nervous systems, innate reflexes, and also the power of human speech, **is a product of evolution.** And for this reason, these structures and abilities are adapted to the real world. The variability and inheritance of talents (language, music), variations of intelligence, mental diseases, different capacities of sense organs (e.g., comparing the performance of eyes),

modifications of anatomical structures in the course of evolution (improvement of the construction of eyes, increase of brain size), point to the existence of selection and of the modification of "intellectual" properties. Many circumstances indicate that our cognition ability is imprinted by the properties of the environment. Examples are the congruence of the construction principles of sense organs seen in groups of animals that are not related, the adaptation of abilities of sense organs and nervous systems to those environmental signals that are essential for the survival of an organism (e.g., ultrasonic hearing in bats, perception of water flow in fishes, three dimensional vision in arboreal animals), the congruence of information we get from different sense organs (if we touch with a hand the thing we see, we find it exactly where we see it; if it has a rough surface, we see *and* feel it). The prerequisite for this evolutionary adaptability, namely the inheritance of these structures and abilities, is a "fact" for biologists. Furthermore, evolutionary epistemology is in harmony with other theories and with results of other sciences (laws of genetics, theory of evolution, the origin of humans, the physiology of sense organs, neurobiology, language theories).

Within the limits of the mesocosmos* relevant for our survival, our cognition apparatus guarantees a realistic image of the environment. This entails the forming of notions through abstractions, an important data processing ability of the nervous system, which in higher metazoans is able to identify invariable parts of patterns in incoming data and to compare them with previous experiences. In this way the pre-scientific process of learning a language leads to the development of words and notions that refer to or represent existing things or properties of things or processes (see Vollmer 1983). However, this primary knowledge is not complete and not in every case a reliable copy of the environment, and thus has to be tested critically and complemented by scientists.

Evolutionary epistemology also refers to results of sensory physiology: it is possible to show that a correlation exists between the stimuli originating in the environment and the reaction of the sense organs and the nervous system. The optimization of this correlation, which partially requires extensive data processing in organisms, has to be interpreted as a result of evolution. Organisms with nervous systems not able to identify those patterns of reality which are relevant for their survival died out, or had to succumb in competition for resources. A gibbon not able to estimate the distance to the next branch correctly falls down when jumping. The central contribution of evolutionary research to epistemology is that the adaptation of the organismic cognition apparatus also includes rational behaviour. The ability to think is also subject to evolution. The cognition ability of organisms is an adaptation to the environment and decides on life and death.

This explains why the knowledge of living organisms corresponds to aspects of reality and why pre-scientific forming of notions is generally reliable (Lorenz 1973, Riedl 1975). (In a wider sense we even can include genetic information used for non-neuronal reactions in what we call "knowledge".)

This, however, does not mean that an experience always leads to a reliable, true description of reality. On the contrary, what is experienced is with good reasons the cause for the formulation of hypotheses. Hypotheses can be revised with the finding of new evidence and thus do not represent absolute truths but rather statements with varying probabilities of being "true". Evolutionary epistemology explains why our perception does not provide an exact copy of our environment: for our survival exact and complete knowledge is not important, but relevant is an appropriate reaction to environmental factors that are useful for the needs and for the reproduction of the individual. This is the motive to call the scientific knowledge of *Homo sapiens* hypothetical, and to call his scientific way to look at nature "hypothetical realism" (Oeser 1987).

* *mesocosmos*: a small part of the universe containing all factors necessary for the survival of an organism.

2. The subject of phylogenetic systematics

First of all, we have to find out whether the items systematists are dealing with are material things, properties of things, processes, or rather mental constructs. The most important of these items are

- properties of organisms,
- the transfer of genetic information between organisms,
- inheritance and modification of genetic information,
- populations of organisms,
- monophyletic groups,
- species,
- speciation processes,
- phylogenetic trees.

It is obvious that the methods used for the acquisition of data and for data processing are also subjects of research in systematics. In this chapter, however, we will focus on terms which probably represent things of the material reality.

The individual organism that lives today or the organisms of former times are the subjects of analyses from which all data on "characters" are obtained. The single organism is of course a real object. The term "character" has been explained in chapter 1.3.7; it can be a material object, a real property of an organism, or a mental construct. The terms "population", "monophylum", "species" and "speciation" will be dealt with in the following chapters. Inheritance and modification of genetic information are the subjects of character analyses (see. ch. 5), which are necessary to gain evidence on historical processes. This approach differs from the mechanistic point of view typical for the study of effects of mutations and modifications of phenotypes found in text books on genetics or developmental biology.

Knowledge on the characteristics of populations, of speciation events (ch. 2.3.1), and also of the processes of inheritance and the modification of genes, is obtained either indirectly, comparing organisms and deducing what happened in the past, or directly, observing processes that occur today. The comparison allows statements on the existence of reproductive barriers, on the chronological sequence of speciation events, or on the genetic distances between species.

The basis for many further conclusions, for example on the evolution of adaptations or on dispersal events, is a phylogenetic tree (= **dendrogram**, cladogram, phylogram, mathematically also tree, tree graph; see ch. 3), which usually is depicted graphically. The inference of trees is a central theme of the following chapters. In combination with further data, for example on the geographic distribution or on the way of life of species, conclusions are possible concerning

- the historical age of a group of organisms,
- the process of evolutionary adaptation to local ecological conditions,
- the existence of radiations,
- the influence of climate changes, continental drift, migrations, the effect of the evolution of other organisms on phylogeny or on evolution of a species.

The study of evolutionary adaptations and of the influence of the environment will not be dealt with in this book. These themes are discussed in the extensive literature on evolutionary biology (which does not mean that systematists should not also be evolutionary biologists). However, the **existence of evolutionary processes** is a fundamental assumption of phylogenetic analyses (ch. 2.7). Although phylogeny can be reconstructed (by analysing whether similar patterns are probably congruent by chance or, alternatively, due to the existence of a common source of information, see. ch. 5.1.1), without this assumption or knowledge of evolutionary biology, it would neither be possible to understand the causes of phylogeny nor to discuss the plausibility of a hypothesis of relationships. Many important questions would not be asked.

2.1 Transfer of genetic information between organisms

An organism that is nearly identical to the corresponding parent organism(s), develops from an undifferentiated cell due to the presence of copies of its parents' or its mother's nucleic acid sequences. This fact justifies the common phrase that "genetic information" has been transmitted (for the term "information" see ch. 1.3.5). This process can take place in different ways, and the following possibilities have to be distinguished:

- horizontal gene transfer,
- clonal reproduction,
- bisexual reproduction.

2.1.1 Horizontal gene transfer

The term "**gene transfer**" refers to the material transmission of nucleic acids from one cell to the next. It is used for the transfer of only a fraction of the genome of the donor organism, a process usually not serving reproduction. In cases in which reproduction takes place, the transmitted genes do not form a major part of the offspring's genome. Receiver and donor generally cannot interbreed and are thus reproductively isolated (a difference to gene introgression by hybridization). The effect is always that an organism shows nearly exactly the genome of the parents (in the sense of a copy of the parents' DNA sequences), and in addition possesses single genes or gene sequences originating from individuals of other populations or species. In prokaryotes, transformation, conjugation or transduction occur regularly. Between eukaryotes, gene transfer is a very rare event, the mechanisms are unknown except for cases where viruses are considered to be the vehicles.

Example: Due to the incongruence between the sequence similarity of "P-elements" of *Drosophila*-species with the supposed phylogeny of these species, it is assumed that the observed similarities between "P-elements" are the result of horizontal gene transfer (Clark et al. 1994). – In Cyanobacteria (and in plastids which developed from endosymbiontic Cyanobacteria) two different proteins occur which are involved in photosynthesis. A protein very similar to one of those is common in green sulphur bacteria and a protein similar to the other one occurs in purple bacteria. The most probable explanation is that due to horizontal gene transfer both genes met in an ancestral cell of Cyanobacteria. – In the stem lineage of Eubakteria, certain ATPases occurring in *Thermus* and *Enterococcus* probably were originally absent. These enzymes correspond to the ATPases of Archaebacteria. The distribution of these genes can be explained with gene transfer (Gogarten 1995). – For fungi it has been shown that species that are not related can transmit linear plasmids after hyphen contact (Kempken 1995). – Gene transfer also occurs between bacteria and fungi (García-Vallve et al. 2000). – Transfer does also exist between endosymbionts and host cells (transposition, see ch. 5.2.2.3).

Horizontal gene transfer can be a source of error in phylogenetic analyses when the distribution of homologies in different organisms does not correspond to the true phylogeny. In the evolution of a population of animals (Metazoa), vertical flow of information from parents to offspring is usually the only mechanism by which genetic information is transmitted.

2.1.2 Clonal reproduction

For systematists the essential characteristic of clonal reproduction is the complete lack of recombination of genes. Facultative or cyclic clonal reproduction as seen in cladocerans, aphids, or cynipid wasps is only a delayed bisexual reproduction and thus is not exclusively clonal. In these species an unisexual and a bisexual generation alternate and recombination does occur regularly, though not in each generation. In clonal populations genetic information is passed on in form of copies of nucleic acid sequences from the parent organisms to the next generation. Clonal reproduction occurs

- in the case of vegetative propagation, and
- in the case of unisexual reproduction (parthenogenesis).

In these systems the path of the "flow of information" (the transmission of copies of DNA) can be visualized in form of a tree (Fig. 20). The stabilizing selection by environmental factors and the limited capacity of habitats are the factors that

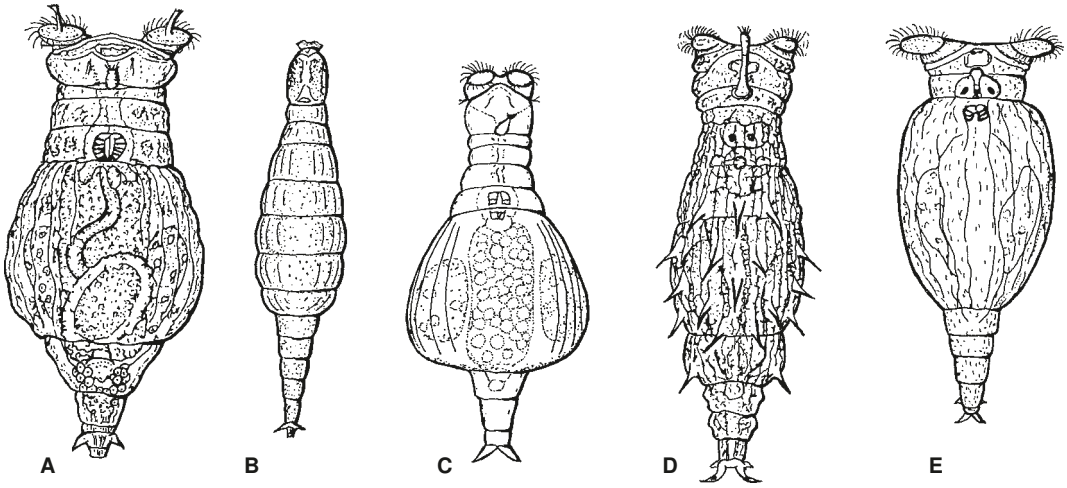


Fig. 17. Bdelloidea (Rotatoria) reproduce exclusively parthenogenetically, whereby a large diversity of forms evolved. **A.** *Mniobia magna*; **B.** *Adineta gracilis*; **C.** *Habrotrocha lata*; **D.** *Dissotrocha aculeata*; **E.** *Philodina megalotrocha* (adapted from Streble & Krauter 1973).

allow the coexistence of only a limited number of very similar individuals, whereas other variants with unfavourable mutations are handicapped and have fewer or no offspring and their genotypes gradually disappear from the population. Such groups of organisms are **clones**, whose genes can always be traced back to one founding individual.

A method for the delimitation of such groups consists in the reconstruction of the origin of a clone (as in the case of monophyletic taxa): all members of the clone must share the same last common parent and therefore possess copies of the same DNA sequences. The genetic similarity of organisms can be used as evidence for the existence of a clone. It is important that there exists no exchange or recombination of genes. Even after several generations, mutations that occurred in a single individual of generation 1 for the first time cannot be found as homologues in offspring of other individuals of generation 1. Some examples for clonal animals: several Nematoda (e.g., species of *Meloidogyne*), Ostracoda (Cyprididae, Darwinulidae), several Phasmatodea (e.g., *Carausius morosus*). Among the Rotatoria the Bdelloidea reproduce exclusively parthenogenetically (Fig. 17). In the Amazon molly *(Poecilia formosa)*, a hybrid species of Middle American representatives of the viviparous Poeciliidae, the eggs only develop after contact with spermatozoans of related species, however without fertilization. The American kilifish *Rivulus marmoratus* reproduces only through self fertilization (Harrington & Kallman 1968). Amongst the angiosperms, apomictic or vegetative populations reproducing exclusively asexually originated several times from bisexual ancestors through hybridization or polyploidy (e.g., in the genera *Dentaria*, *Mentha*, *Acorus*, *Potentilla*, *Taraxacum*, many Rosaceae, Cichoriaceae, Poaceae). (Apomixis is in botany the development of embryos through parthenogenesis or from vegetative cells).

2.1.3 Bisexual reproduction

The flow of genetic information in groups of organisms with normal sexual reproduction can be illustrated in form of a network (Fig. 20): the genetic information present in an individual cannot be traced back to only a single ancestor. The whole group of individuals being part of such a network and living at the same time form either a **potential** or a **functional reproductive community**. Only if the individuals of different subgroups can encounter each other in nature and can reproduce successfully the group is a *functional* unit. On the other hand, a *potential* reproductive community consists of populations which cannot meet and interbreed in nature, but contain individuals which can hybridize successful-

ly in an experiment. Such populations are delimited from others not belonging to the same reproductive community because gene transfer between different reproductive communities is not even possible in experimental crossings. Functional reproductive communities form natural, material systems (see. ch. 1.3.2), potential reproductive communities, however, are not material systems. The sum of the genetic information or of all variants of genes of such a system is called the **gene pool**. This word is, of course, only a metaphoric expression, a material pool does not exist. The composition of the gene pool, for example the presence and frequency of alleles, can change continuously in the course of time, without loss of the delimitation to other systems.

The term "**biopopulation**" is sometimes used as synonym for a species concept (Mayr 1963, 1982, Mayr & Ashlock 1991) or for groups of organisms of a species in the sense of a potential reproductive community (Mahner & Bunge 1997). More on species concepts in chapter 2.3.

2.1.4 The special case of endosymbionts which evolved to organelles (mitochondria and plastids)

Obligate intracellular endosymbionts which never leave their hosts and reproduce only asexually are clones whose evolution depends on their hosts' survival and evolution. The transmission of genetic information from one generation to the next usually takes place in the same way as in free-living clones, even when the host reproduces sexually. Not all endosymbionts are as highly specialized as some organelles, there exist all transitions in nature beginning with loose associations between different species (Margulis 1993):

- facultative endosymbionts which in nature occur also outside their hosts (e.g., bacteria living in *Paramecium*, *Rhizobium* in Leguminosae, *Chlorella* in *Paramecium bursaria* or in *Hydra*),
- obligate endosymbionts (e.g., cyanobacteria in *Glaucocystis*, methanogene bacteria in *Pelomyxa*), and
- endosymbionts depending on the host's nuclear genome, because the symbiont's own genome is not complete: the symbiont evolved as functional unit of the host (e.g., cyanelles in eukaryote unicellular organisms, in Cnidaria, "plastids" of *Euglena*). A transfer of genes from the endosymbiont into the nucleus of the host cell is possible (transposition, ch. 5.2.2.3).

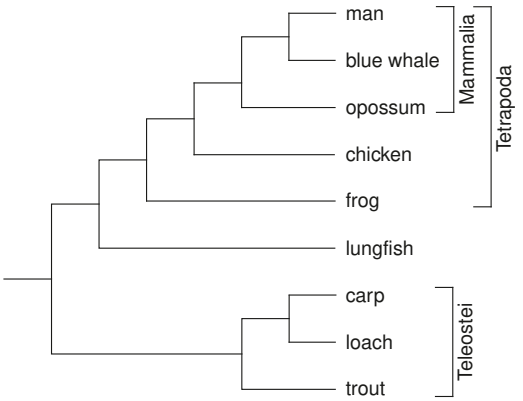


Fig. 18. Phylogenetic relationships of some vertebrates, calculated from complete mitochondrial genomes (mtDNA-sequences; after Zardoya & Meyer 1996).

Mitochondria and plastids belong to the latter group of endosymbionts which form a functional unit with their host cell. We must expect that genetic drift and selection of mutants of endosymbionts depend, among others, of the mode of life and of the population dynamics of the host organisms whenever the symbionts are no longer able to leave their hosts. Therefore, the genomes of the symbionts should evolve parallel to those of the host species as long as no gene exchange takes place between endosymbionts of different hosts (Fig. 18). This is true for mitochondria and plastids, organelles that have their own genomes, as well as for cyanelles and other endosymbiontic protists which cannot leave their hosts any more. If we assume that the substitution rate of nucleic acids depends on the number of replication events, then symbionts and hosts should show different rates (see Moran et al. 1995), because the symbionts can multiply independently of the hosts' reproductive cycle.

It is well established that the mitochondrial genome of animals in most cases is inherited through the maternal germ line. Male mitochondria present in sperm cells are not transported into the zygote. Because of this phenomenon, clones of mitochondria evolved that are only transmitted maternal-

ly. Therefore, mitochondrial genes are very useful for the inference of the phylogeny of the carriers of these organelles as well as for studies of population history, for example, to analyse the dispersal routes of female specimens. There are only few exceptions to the rule of maternal inheritance: it is known for mussels that there exist sex-specific mitochondrial DNAs (Geller 1994, Liu & Mitton 1996).

In plants, "foreign" plastids can get into a species through hybridization. The nuclear genome of theses species may then show a different phylogeny than the plastids. Results of phylogenetic analyses of some Californian Asteraceae indicate that hybridization of the annual *Microseris douglasii* with *M. bigelovii* produced individuals that possess most of the nuclear genome of *M. bigelovii* (the maternal nuclear genome of *Microseris douglasii* was eliminated) and plastids of *M. douglasii*, whereas in some cases the chloroplast of *M. bigelovii* was retained (Roelofs & Bachmann 1997). In such cases the phylogeny reconstructed for the plastids differs from that of the nuclear genome.

2.2 The population

The term "population" is used in biology for an accumulation of organisms implying either the assumption that they belong to the same species or that they show properties which often are also thought to be characteristic for a species (e.g., genetic similarity, potential for cross fertilization of individuals). The term may be used for very different groups of individuals (Mahner & Bunge 1997):

- an accidental assemblage of individuals,
- individuals living in a limited place (a valley, a pond),
- all individuals living at the same time or all individuals of a species viewed in space and time (depending on the species concept!),
- all members of a clone,
- all individuals between which gene flow is principally possible (a potential reproductive community or "biopopulation"; for example all living horses),
- all individuals with a network of real reproductive contacts (a functional reproductive community; e.g., all horses of a herd).

To get an objective classification of groups of organisms which is relevant for biological sciences, we have to search for laws of nature. Lawful phylogenetic or genealogical relationships exist only between

- all members of a **clone**, because they originated as offspring of a single individual through vegetative or unisexual processes and thus are genetically nearly identical,
- the members of a **functional reproductive community**, which share many genes and mutations due to sexual processes and form a natural, material system,
- the members of a **potential reproductive community** (biopopulation), because they are the descendants of the same ancestral population, sometimes of only a single ancestral pair, and thus are genetically very similar to each other. In this group sexual contacts do not necessarily have to be realized, for example when populations are separated by insuperable physical barriers.
- For similar reasons, all descendants of a single ancestor organism or of a pair of organisms are lawfully related, even if they are not considered to be members of the same species. Such groups are called **monophyla** (see also ch. 2.6).

The members of a clone or of a *functional* reproductive community have **tokogenetic relationships** within the populations (parent-child-relationships). These are lacking in the *potential* reproductive community.

Identification of group members

A species concept is not necessary to identify the members of a reproductive community, because the members can be identified on the basis of successful mating, they are part of a network of reproductive relationships. However, members of a single clone cannot be identified on the basis of their reproductive activities, except in those cases where the origin of the whole clone could be monitored (e.g., in experimental cultivations

of bacteria). Indirect evidence for membership is the great similarity of genes in different individuals that can be explained assuming descent from only one ancestral individual. This hypothesis can be verified with an analysis of genetic distances, which should show clusters of individuals, each cluster representing a single clone, or uncovering unique apomorphies characteristic for a single clone. A similar case are the potential reproductive communities, whose members are usually grouped due to morphological or genetic similarities. This procedure is typical for the species concept used by laymen. Biologists also evaluate the visible similarity as evidence for descent from a common ancestral population, however, in some cases similarity is not the best criterion (see chapter 2.3).

At this point we do not need the terms "**biospecies**" for potential reproductive communities and "**agamospecies**" for clonal populations, because they require the existence of a species concept. To study the existence of natural processes that are the basis for the above-mentioned groupings a species concept is not needed. Mahner & Bunge (1997) use the word biospecies for the biological species, other authors just mean the reproductive community (s. Sudhaus & Rehfeld 1992). It is more important in this context to study the phenomena that are observable in nature than to discuss definitions.

Organism: an individual living being (living object). (Note: a fossil is not an organism, but what remains of it).

Clonal population: a group of organisms with nearly identical genes, originating from a single ancestral organism through vegetative or unisexual reproduction.

Potential reproductive community: a group of spatially separated organisms living at the same time, whose members can mate successfully with partners of the same group in an experiment. Even though they may be spatially and thus reproductively isolated, they are assigned to the same group.

Functional reproductive community: group of organisms living at the same time, able to reproduce successfully with partners of the same group. There are no barriers to gene flow in this group.

The ontological status of populations

Of these groups only the functional reproductive communities are **natural material systems** (see below). The others, however, are **constructs**, which proved to be useful in ecology and evolutionary research, because these groups consist of genetically identical or very similar individuals. Since in the case of clones and potential reproductive communities the group members are actually closely related, these mental constructs are at the same time **natural kinds** with a relation to reality that can be objectively identified, even though a material object "group" or a material system is not present.

Natural material systems and individuals exist only in the actual time horizon ("now"; ch. 1.3.2). Therefore all considerations on populations or on parts of phylogenetic trees viewed in a four-dimensional space-time-framework are constructs.

Often parthenogenetic populations emerge from bisexual ones. Within the Philosciidae (terrestrial isopods, woodlice) there are bisexual as well as parthenogenetic species (Johnson 1986); many freshwater-ostracods, and many species of collembolans are parthenogenetic (Danielopol 1977, Palevody 1969), arctic populations of *Daphnia pulex* are exclusively parthenogenetic (van Raay & Crease 1995).

As already mentioned, functional reproductive communities can be considered to be material systems. **Clones** are systems only if contemporary individuals mutually influence each other, so that the community acquires new properties which are absent in single individuals, for example when they transform their environment and thus create better conditions suitable for all members of the clone. Clones and bisexual reproductive communities are not similar in respect to the presence of system relationships: between the members of a clone gene flow does not form a network (see Fig. 20). If there is no other factor influencing the development of a uniparental population, such populations do not develop as a single unit: the genes present in a clonal population do always stem from a single individual (unless horizontal gene transfer occurs: ch. 2.1.1) and the ancestor-descendant-lines diverge without genetic feedback from other members of the population. **Common descent is not a process occurring now and not a system property.**

In bisexual organisms, the stock of genes of an individual has its roots in a large number of ancestors. These

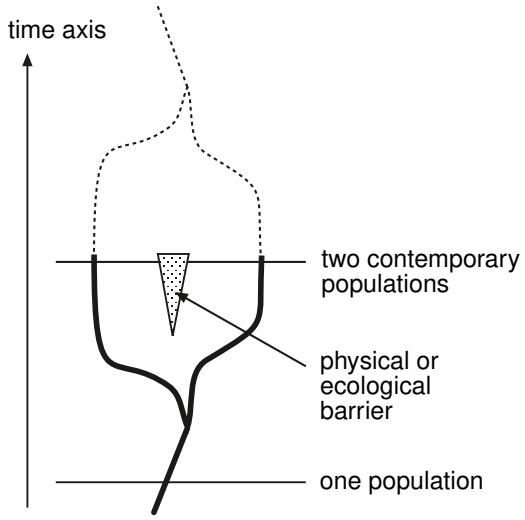


Fig. 19. Scheme for the possible future development of 2 separated populations ("populations" in the sense of functional reproductive communities).

numbers are not only high by summing ancestors up in time along a stemline, but also due to many individuals that lived simultaneously. Genetic information is distributed in a **functional reproductive community** through recombination. Therefore, such a population functions like a superorganism, whose parts are not physically joined and can be exchanged like cells in a body. The "gene flow" provides for the information (genetic "instructions") needed by the replacement (a new individual); the information originates from other parts of the organism. Whenever a netlike gene flow is detectable in animals or plants, we can consider the organisms of the corresponding population as being members of the same species (see species concepts in chapter 2.3).

The **size of populations**, meaning the number of organisms in question, apparently depends on the interval of time under consideration. The number of considered humans is larger in the interval from 1.1.1930-1.1.1970 than at the 1.1.1970. Both groups comprise other sets and thus are not identical. Even if one forgets that a material system only exists in the present point in time, discrepancies rarely occur in statements about populations: the real population does not exist yesterday *and* today, but only now. What we visualize with an age pyramid or a length-frequency diagram is the result of an analysis and not "the population". The phrase "in the population of Darwin finches the average size of the bill length changed in the course of 2 years" is nevertheless understood unambiguously by biologists: the frequency of alleles influencing bill length changed in successive time levels. (One should also be aware of the fact, that the group "population" does not have a property "length of bill", but only the individual bird shows a real property.)

Divergence of populations: the following consideration shows that we are dealing with constructs most of the time when we talk about the history of populations: two populations of a potential reproductive community can live absolutely separated for a while, but later can join again so that over a longer period of time only *one* population can be identified (Fig. 19). Should the populations not be able to merge again physically and if they diverge genetically until successful mating between individuals of the populations is not possible any more, with the evolution of a reproductive isolation of the subpopulations the original single material system stops to exist as a unit (see the analogy of a bush fire that divides into two, ch. 1.3.2). This means that the same fact existing today (two physically separated populations) could lead to different groupings and material systems in the future depending on the further development. As the future of two spatially separated reproductive communities cannot be known, it is futile to debate whether at this point both populations are still a "unit of nature" or not. At present they do not form a natural material system, but they could become one in the future, if they get into contact. But they could also diverge for ever.

These considerations have consequences for the status of populations as part of a biological "species": if it can be predicted with certainty that the separated populations will merge in future (Fig. 19) all living members of spatially separated potential reproductive communities should be considered to be members of the same biological "species". However, exactly the same organisms would be assigned to two separate species if the future divergence of the populations is irreversible.

Since the status of populations as members of one or more species depends on predictions about the future development of the populations, this example should help to elucidate that the ontology of the "biological species" is generally viewed wrongly: the species is not a "unit of nature" but rather a construct (see ch. 2.3).

The divergence of populations can be established objectively: if the genetic distances of all pairs of individuals within and between populations fall into two clusters (see ch. 8.2), the average distance between the two clusters can be calculated. If this distance increases with time the populations diverge. Gene flow between members of these clusters can reduce the distance.

Demarcation of populations: In nature, all individuals of bisexual species classified by us as members of the same group can have contact through different biological interaction (competition, mating, cooperation). For a **genetic classification**, however, only the consequences of reproductive contacts are of interest. Horizontally, i.e. at a given moment of time, a functional reproductive community is separated from others through absence of system relationships (see term "system" in ch. 1.3.2). On the other hand, members of clones and of potential reproductive communities are classified due to their genetic similarity, while the possibility to test for cross-breeding experimentally is only rarely used. **Along the time axis** there are **no natural limits** for reproductive communities nor for clonal populations except extinction. They diverge or, viewed backwards in time, fuse gradually with each other (Fig. 19, 20, 28). From the first eukaryotic cell to humans there was an uninterrupted succession of generations descending from each other. The lack of limits in the vertical axis, i.e. in time, is normal for natural systems (see ch. 1.3.2).

The **dimension of time** is not excluded from descriptions of populations: it is customary in biology to talk about "population growth", "seasonal population changes" etc. However, to name different stages of these systems at different time levels is difficult because, viewed realistically, individual populations currently assigned to different species certainly differ today (in the horizontal plane). The differences (e.g., genetic distances) can be measured objectively, but, in the vertical dimension (along the time axis) a sharp boundary between groups cannot be found (e.g., in Fig. 20 between group **A** and group **X** or between **D** and **Y**). Each attempt to define a boundary along the time axis is an arbitrary act, in nature this boundary does not exist. Therefore, with the naming of a population a time interval for which the name will be valid has to be defined clearly. This observation is of great importance for the discussion of species concepts.

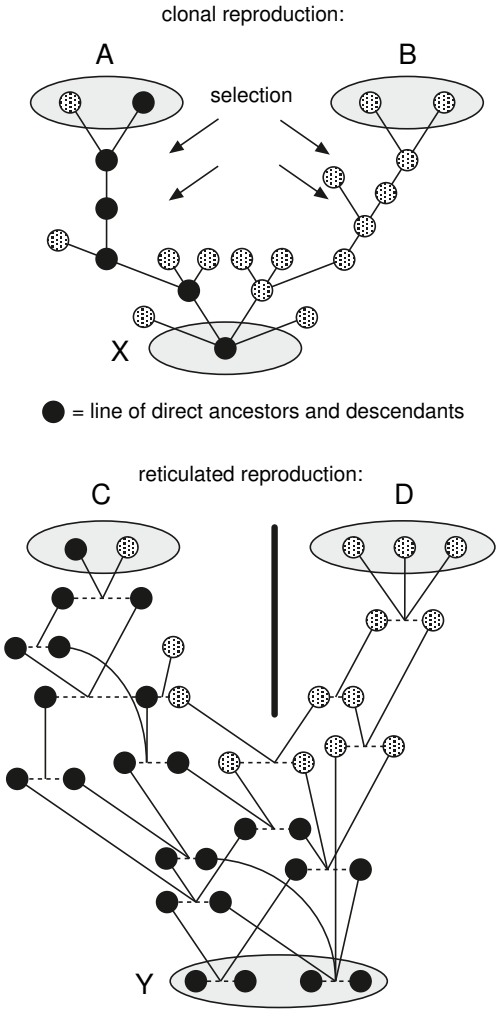


Fig. 20. Clonal and bisexual populations have different system properties: in the upper graph each individual has only one parent and one ancestor in each ancestral generation, in the lower illustration, however, the number of ancestors (black circles) increases in each generation while going back in time. The vertical bar symbolizes a physical or reproductive barrier. The circles represent individual adult organisms, the lines represent descent. The horizontal distance symbolizes the extent of the genetic difference, the vertical axis the distance in time.

Origin of new groups: groups have to diverge genetically to become distinguishable. The origin of several simultaneously existing and distin-

guishable groups of clonal organisms (A-D in Fig. 20) from a uniform stem population has to be explained with the occurrence of mutations, extinction of specific mutants due to selection, and the preservation of identical gene copies within a group. Examples: the Caucasian lizard species *Lacerta valentini* includes parthenogenetic populations, whose females produce triploid, sterile descendants after mating with males. A gene exchange with bisexual races is not possible any more although the animals are morphologically very similar to each other and are assigned to the same "species". A further modification of the parthenogenetic population could tempt people to give it a new species name. – In North America, parthenogenetic "races" of *Daphnia pulex* evolved into diploid and polyploid populations which can be distinguished morphologically. – Parthenogenetic populations of Spanish brine shrimps *(Artemia parthenogenetica)* differ morphologically and genetically to such a degree that they could be endowed with a separate species name (Perez et al. 1994). – Within the Rotatoria, the Bdelloidea are exclusively parthenogenetic, nevertheless they evolved into a variety of life forms for which different species names are in use (Fig. 17).

On the other hand, in bisexual organisms the prevention of reproduction between genetically more or less similar individuals, i.e. the evolution of a "reproductive barrier", is necessary to interrupt the continuum of genetic variation within a stem group (Y in Fig. 20).

The term "**reproductive barrier"** has to be explained briefly. This is of course only a metaphor for the effect of very different processes which lead to the same result: starting with a uniform population at least two new groups originate in time in which group members can produce offspring only with members of the same group. For a final separation of the populations it is important that the "barrier" has some genetically fixed causes within the organisms. The terms "inborn reproductive barrier" and "reproductive isolation" refer to the fact that some organisms possess the features necessary for successful mating within a population, whereas these are lacking in other organisms. Thus not the "non-properties", but the existence of real novelties (mechanical structures for copulation, pheromones, courtship behaviour, receptor molecules) is paraphrased with these terms

2.3 The "biological species"

"When the views entertained in this volume on the origin of species, or when analogous views are generally admitted, we can dimly foresee that there will be a considerable revolution in natural history. **Systematists will be able to pursue their labours as at present**; but they will not be incessantly haunted by the shadowy doubt whether this or that form be in essence a species. This I feel sure, and I speak after experience, will be no slight relief. The endless disputes whether or not some fifty species of British brambles are true species will cease. **Systematists will have only to decide (not that this will be easy) whether any form be** sufficiently constant and **distinct from other forms**, to be capable of **definition**; and if definable, whether the differences be sufficiently important to deserve a specific name."

(Darwin 1859)

We have to differentiate
- the question concerning the reality of phenomena or of objects in nature we want to name with the term "species" (the theoretical problem), and
- the criteria for the identification of species (the practical problem).

The solution of the practical problem depends on the definition of the term "species" we want to use. In the following paragraphs it will be explained why evidence for the existence of a process we want to call "irreversible divergence of populations" has to be presented as criterion for the delimitation of "species".

The question has to be asked, whether in nature there exists a material thing or system that can be called "biological species". From the colloquial usage of the term "species" it follows that always organisms are meant,

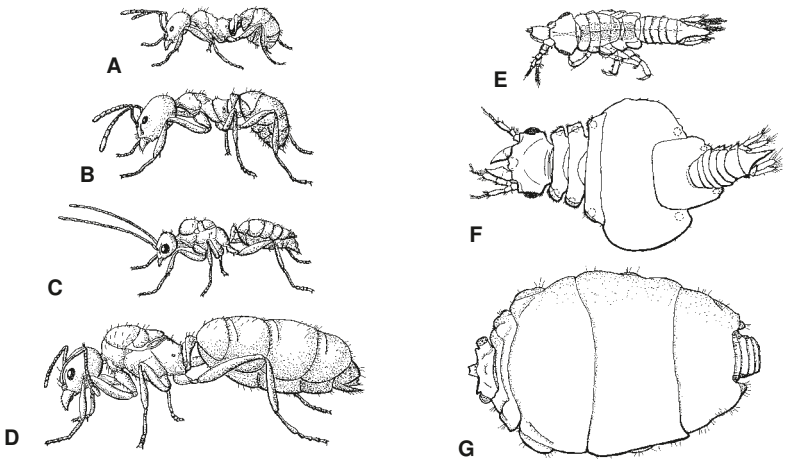


Fig. 21. Morphological variations within a species. **A-C**: casts of the ant *Aneuretus simoni*: workers (**A**), soldiers (**B**), males (**C**) and females (**D**). Stages of the marine isopod *Caecognathia calva*: larvae (**E**), males (**F**), females (**G**). A-D after Wilson et al. 1956, E-G after Wägele 1987.

a) which through the process of sexual reproduction form a material system reproductively isolated from other such systems.
b) When the ability for cross-breeding has not been observed, the individuals are assigned to the same species due to their morphological or genetic similarity. In this case the assumption that similarity is generally an evidence for species membership is based on analogous phenomena observed in reproductive communities where successful mating has been monitored. Similarities are also used for the classification of clonal organisms.

Biologists have discussed the definition of the "biological species" so controversially that this fact alone raises the suspicion that in nature a material thing or system which can be discerned objectively as "species" from the surroundings does not exist (e.g., Bachmann 1998). This contrasts to the opinion of several authors who see a "species" as a real evolving "individual" (e.g., Ghiselin 1974). The material existence of living organisms and of populations "as material systems" cannot be questioned (see ch. 2.2). Furthermore it can be observed that organisms reproduce and that the descendants are morphologically and genetically very similar to their parents.

Similarity can be measured or estimated objectively (e.g., length of legs, chemical composition of secretions, frequencies of sound in songs, number of bristles on an antenna). Therefore, individual organisms can be grouped objectively such that they are more similar within a group than between groups. Small groups whose members are similar to each other are traditionally called species, but larger groups are also recognizable (ungulates, birds, sharks), as well as even smaller ones ("races"), which can be identified as subgroups of species on the basis of observed cross-breeding. Difficulties arise because groups of organisms which are well differentiated externally and are assigned to different species (tiger in comparison to lion) may nevertheless mate successfully. On the other hand there are organisms which cannot be differentiated externally but nevertheless belong to different populations separated by reproductive barriers. Examples: cryptic species such as *Anopheles gambiae* and *Anopheles arabiensis* (Culicidae) can be identified with RAPD-markers (Wilkerson et al. 1993) but are difficult to distinguish morphologically; cryptic species of mussels were identified with enzyme electrophoresis (McDonald et al. 1991); cryptic species of corals of the genus *Montastraea* (Knowlton et al. 1992) could only be discerned genetically. Another phenomenon which has to be taken into consideration is intraspecific polymorphism (Fig. 21).

Similarity is not a criterion of universal utility that can be applied to distinguish "species" (Fig. 21), neither at the level of morphological charac-

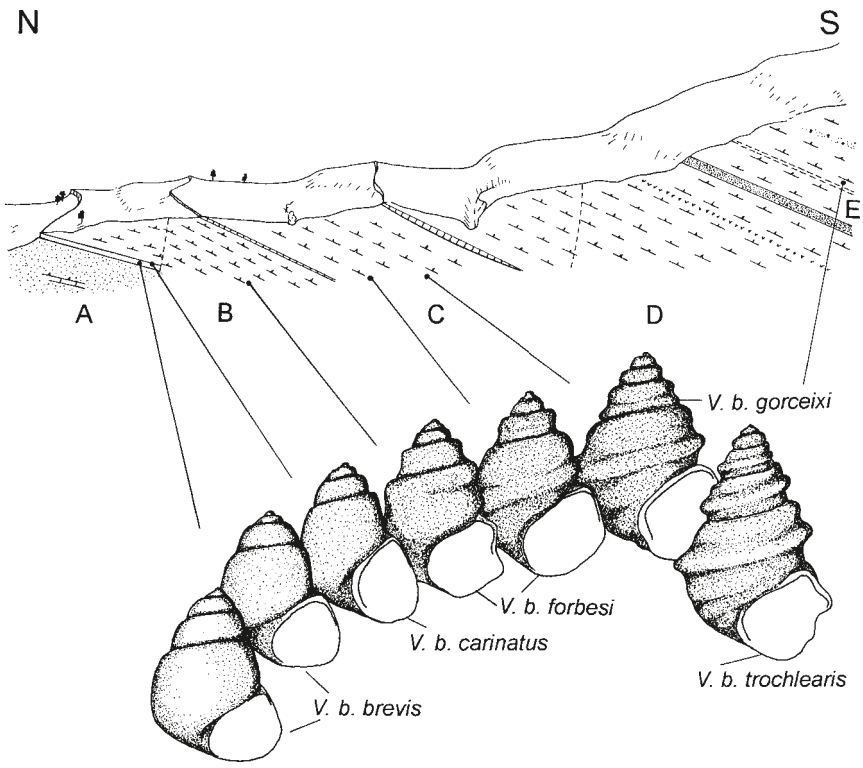


Fig. 22. Morphological variation of the fossil fresh water snail *Viviparus brevis* during the Pliocene and Pleistocene of the island Kos (Greece). Species boundaries cannot be defined objectively as long as evidence for the irreversible divergence of another line of populations is not found (from Willmann 1985).

ters nor with the help of genetic distances. Populations are not constant, they develop in the course of time. Morphology of individuals changes during their ontogenesis, morphology as well as gene sequences vary between individuals, some populations change faster than others. Examples: sister species within mammals have larger distances of the cytochrome b gene than in birds or fish (Johns & Avise 1998). – Fossil populations of fresh water snails of the Greek island Kos show an unusual morphologic variability through time (Fig. 22) which is correlated with environmental changes.

Species which are similar to each other often can be characterized by their differences in use of resources and in habitat requirements. For example, *Eucalyptus*-species in Southeast Australia have different temperature preferences; larvae of cerambycid beetles use different host plants (e.g., *Pogonocherus fasciculatus* lives in dry, small branches of coniferous trees, *Pogonocherus hispidus* in dead branches of deciduous trees); on European rocky shores the barnacle *Chthamalus montagui* occurs slightly higher up in the marine supralittoral than *Chthamalus stellatus*. A metaphorical expression is that "species" occupy their own "ecological niches". These differences, however, are not reliable indicators for the differentiation of species, because there also exist morphologically adapted races with different preferences (for example, fresh water races of the amphipod crustacean *Gammarus duebeni* have larger kidneys than in populations living in brackish water; races of subterraneous woodlice living in caves are depigmented, etc.). But, there are also species with nearly identical habitat requirements which cannot cross breed (e.g., the littoral snails *Hydrobia ulvae* and *Hydrobia ventrosa*, which only feed on particles of different size when they live in sympatry).

There exists no objective criterion to differentiate two populations succeeding each other consecu-

tively in time (typically each represented by a few fossils), and to assign them the status of a species. Such "**chronospecies**" would only be distinguished according to the extent of morphological change. In this case, the definition of species boundaries results from purely subjective decisions. To sort out species, the presentation of evidence for the existence of a "reproductive barrier" is not suitable either, because there also exist clonal groups of organisms where a reproductive barrier isolates *all* individuals. If the species concept shall be applicable to all groups of organisms, it also has to include the term "**biospecies**" (for bisexual populations) as well as the term "**agamospecies**" (for clonal populations). Not all definitions used for biological species fulfil this condition.

Biospecies: term referring to a sequence of parent and offspring generations of a potential reproductive community. The term is used only for bisexual populations. The species begins with the genetic divergence that separates it from a sister species (a "speciation event") and ends either with extinction or with the next "speciation".

Agamospecies: term used for a group of related clonal organisms which diverges genetically from other groups of organisms.

Chronospecies: a group of organisms which are considered to be conspecific and lived in a defined period of time. Different species names are applied to morphologically distinguishable organisms (usually fossils) that existed in different periods of time. The species is not necessarily delimited by speciation events.

Some definitions of the biological species

Definition	Problems	Author
Species are groups of organisms with the same morphology (*morphospecies*; typological species concept).	An objective distinction between races and species is not possible.	Linnaeus 1758
Species can be discerned from varieties by the existence of intermediate forms within the species and by the different extent of morphological variations within and between species; species have some specific constant characters.	Boundaries between species along the time axis are not considered, the degree of dissimilarity necessary to propose a species status is chosen subjectively.	among others: Darwin 1859 (see Grant 1994)
Species are groups of interbreeding natural populations that are reproductively isolated from other such groups (*biological species concept, biospecies*).	The limits in time are not considered, clonal populations are excluded.	Mayr 1942, 1969
A species is a lineage of clones or of ancestor-descendant-populations occupying an adaptive zone minimally different from that of any other lineage in its range and which evolves separately from all lineages outside its range (*ecological species concept*).	Lineages cannot evolve; the "adaptive zone" is a concept that cannot easily be related to empirical data; the definition is not applicable when populations of two species use the same resources in the same area, coexisting for some time in displacement competition; local races also fulfil the requirements	Van Valen 1976
Species are reproductively isolated groups of natural populations. They originate through a speciation event and end with the next speciation or vanish through extinction (*phylogenetic species concept*).	Clones are excluded. This concept depends on the definition of the "speciation event".	Hennig 1982

An evolutionary species is a single lineage of ancestor-descendant populations which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate (*evolutionary species concept*).	A "lineage" viewed within a four-dimensional space-time-frame cannot evolve. It remains unsettled when a lineage starts and when it ends. Evolution also occurs at the level of local populations.	Wiley 1978, 1980
A species is the most inclusive group of bisexual organisms having the same reproductive system (*recognition species concept*).	Clones are excluded; species limits along the time axis are not defined; different organisms which hybridize producing infertile offspring are included in the same species.	Paterson 1985
A species is a cluster of organisms that cannot be subdivided further, with ancestor-descendant relations and with diagnostic characters lacking in other clusters.	Diagnostic characters are also found in races and in larger groups of species (higher ranking taxa); the distinction between races and species is not possible. The identification of "diagnostic characters" depends on the scientist and is not a property of the organisms.	Cracraft 1987; similar in Mishler & Brandon 1987
A species is the most inclusive group of organisms which have the potential for genetic and/or demographic exchange *(cohesion species concept)*.	This does only apply for bisexual organisms.	Templeton 1989
A species is a biospecies if, and only if (i) it is a natural kind and (ii) all of its members are organisms (present, past, or future). The "species as natural kind" is a group of material objects with the same lawfully related properties.	Problems exist in the determination of relevant "properties", the assignment of morphs and varieties to the same species and the delimitation of the species in time.	Mahner 1993, Mahner & Bunge 1997
The species concept refers to a group of ancestors and their descendants, which diverge irreversibly from other such groups along the time axis. The species does not contain irreversibly diverging subgroups (*phylogenetic species concept*).	For populations that are currently physically isolated, it can often not be predicted whether they will diverge irreversibly in future.	–

Some of these definitions only explain how species are recognized and do not state what species are. However, a general agreement exists that a species consists of organisms that have an ancestor-descendant relationship with each other in a four-dimensional space due to reproduction, and that these organisms may furthermore either have the potential to mate with each other (bisexual organisms) or at a given time level they share genes that are nearly identical (in clonal organisms). For further comments on species concepts see Wheeler & Meier 2000.

The last definition of the biological species listed in the table is favoured in this book.

2.3.1 The species concept as a tool of phylogenetics

The **term** "**species**" names a **construct** (a logical class) used to classify some of the properties of organisms and aspects of their history, depending on the personal attitude of the scientist (see definitions of the "biological species" in the preceding paragraph). Terms naming logical classes have to be defined exactly if they shall serve scientific communication because they do not refer to singular processes, material objects or systems. To define a term that can be used unambiguously, it is necessary that it refers to a material or intellectual entity. A species is neither a specific material object, nor in each case a mate-

rial system (see term "system" in ch. 1.3.2). The term is used for clonal populations and for bisexual populations as well. Both types of populations do not show natural boundaries in time ("vertically") except extinction events.

As already discussed, clones and sexual reproductive communities can be distinguished objectively from other clones or reproductive communities horizontally (at a given time level), however, not vertically (in the dimension of time), because there exist gradual transitions between two species (Fig. 20). However, this vertical delimitation is necessary to allow the naming of groups of organisms, a prerequisite for an unambiguous communication between scientists.

Biologists have to admit that the vertical delimitation is done arbitrarily, but according to a pragmatic point of view, referring to historical processes that can be inferred and to properties that can be discovered; the resulting hypotheses can be tested intersubjectively. The following observations can be used for the distinction of species:

- the occurrence of **new characters**,
- a **genetic distance** that exceeds a given limit (see definition of "genetic distance": ch. 8.2),
- a lasting interruption of gene flow (the origin of a **reproductive barrier**),
- the **genetic divergence** of populations.

Which of these observations are used to discern species depends, among others, on the species concept used. The phylogenetic species concept requires the evidence that populations show an irreversible genetic divergence in relation to other such groups.

The term **genetic divergence** has to be explained: it represents the observation that the gene pools of two populations develop in different directions through accumulation of mutations and often at a different speed (with different substitution rates). Objectively measurable parameters are, among others, genetic distances and discrete genetic or morphological differences, which remain smaller between individuals within a population than between different populations. The distances between all organisms result in separated clusters that can be visualized graphically. The reasons for the genetic similarity within a functional reproductive community are (a) the selection pressure, which can have a variety of effects, for example on shapes (e.g., different shapes of bills of Darwin finches) with the result that the carrier of optimal variations will reproduce more successfully, and (b) the sexual combination of genetic information which is passed on to descendants. This "gene flow" between groups and generations of organisms (e.g., between herds of giraffes) prevents on the long run that the divergence of local herds, swarms or clans becomes irreversible. The genetic divergence and diversification of clonal organisms is only determined by selection (ch. 2.2).

The vertical axis in Fig. 23A represents time, the horizontal axis genetic distances (e.g., measured as the number of sequence differences). Due to the fact that in nature new mutations or characters occur with high frequency, nearly an unlimited number of boundaries between species could be determined if characters are selected arbitrarily to define a species and if species were differentiated only by evolutionary novelties. The same holds for genetic distances. Therefore, the sole objective **criteria** that remain are the **genetic divergence** and the **ability to reproduce**. Distances and discrete characters can only serve as more or less reliable **evidence** that these criteria are fulfilled.

In **clonal populations** only the genetic divergence can be seen or measured. Only when two groups of individuals can be clearly differentiated due to their visible or measurable characters, there is reason to assume that they originate from different ancestral individuals and are subject to different selective forces.

In **bisexual** populations genetic divergence and loss of the capacity to interbreed are linked processes. A strong divergence, either boosted by selection or by random genetic drift in isolation, can lead to a loss of those properties necessary for a successful mating with individuals of a sister population. The divergence progresses further after appearance of the reproductive isolation of the sister populations. **Irreversible divergence** does only exist when the gene flow between the diverging populations is interrupted once and for all.

In practice, it proves to be useful to discriminate groups of organisms each showing the following properties:

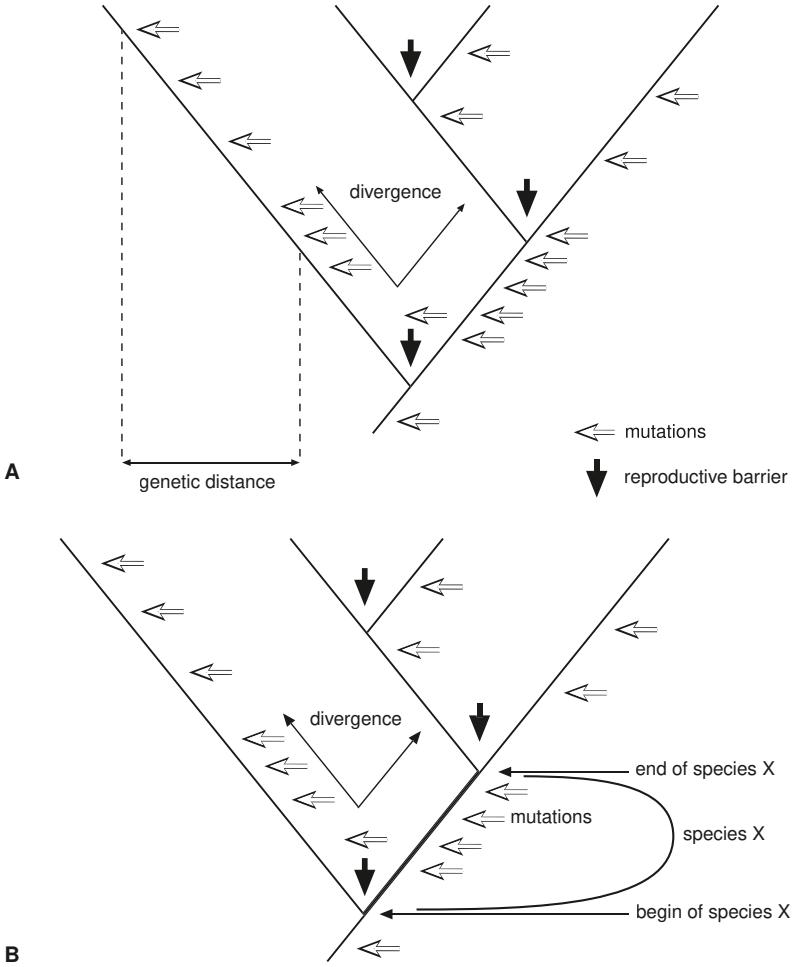


Fig. 23. A. "Speciation events" are processes resulting in the irreversible genetic divergence of populations. In this figure, the genetic distance is represented simplistically in only two dimensions; in reality, the lineages diverge in a multidimensional space. (Attention: the "genetic distance" as used here does not refer to the distance between pairs of single individuals; it is the "generalized distance" between groups of individuals; see ch. 8.5). **B.** Coverage for a taxon of the category "species".

- high congruence of genetically determined characters,
- congruence of ecological requirements,
- descent from the same ancestral population,
- the same changes in the genome of individual organisms during the course of time (also called "historical development" of a population),
- sexual recombination of genes is only possible within the same group (case of the functional reproductive community).

Therefore it is justified to state that the most useful species concept is the one which refers to the series of ancestors and descendants situated between irreversible divergence events (Fig. 23B). Metaphorically, these are the branches of a phylogenetic tree between two nodes, each node representing an irreversible divergence event. It is convenient to define the beginning of a species with an irreversible divergence, and the end with the next one. It would be circular to define the species with the speciation event. Therefore the term "divergence event" is more suitable, it describes the effect of processes observed during

"speciation". This concept corresponds to the **phylogenetic species concept** of Hennig (1982), however avoiding the term "speciation".

The phylogenetic species concept implies that after a divergence event (the "speciation") both descendant populations each are by definition members of a new species; they are distinguished from the mother population even when one daughter species is genetically nearly identical to the mother population. The mother population represents the historically youngest part of the "stem species". The "**survival of the stem-species**" only depends on the definition of "species". If species A splits into species B and C, A and B may have the same phenotype (they are the same morphospecies), but A is the phylogenetic ancestor species of B. We are using here a definition because, as we have seen before, natural boundaries do not exist and the species concept is needed as a tool in biological sciences. By the way, the question if a stem species survives or not is absurd if taken literally, because species do not live (see below).

The phylogenetic relationships in our mentally reconstructed phylogenetic tree are different after the speciation (after the end of species X in Fig. 23B) than before (species X in Fig. 23B) and independent of the extent of the genetic divergence at a specific time horizon.

The following findings concerning the species concept are important:

- The species as "entity of nature" in the sense of a material system can only be a single functional reproductive community. This does rarely apply in the real world.
- Species often are not "entities of nature" but they sometimes consist of clones, more often they are composed of several functional reproductive communities.
- The definition of the species concept ("Which species concept do we need in biology?") is a convention.
- It is useful to define the beginning and the end of a species with irreversible divergence events, because in this way in a dendrogram that represents the real phylogeny single branches can be objectively identified and named, even when one of the new species is genetically nearly identical to preceding one.
- Thus the coverage of a species name ends where at least two new species begin.
- In bisexual populations the origin of a lasting "reproductive barrier" (ch. 2.2) is linked with an irreversible divergence. In clonal organisms the divergence is produced by selective factors controlling the "direction of evolution". The set of processes that produce the initial divergence is also called "**speciation**".
- In bisexual organisms the species marks the boundary between tokogenetic and phylogenetic relationships: where two populations diverge irreversibly, reproduction between individuals of sister populations ceases, but (metaphorically) two new stem lineages of future taxa originate. (Explanation: a tokogenetic relationship is the one due to descent of a child from its parent(s)).
- The basis for the discrimination of species have to be either reconstructed phylogenetic trees or evidence for the irreversibility of the genetic divergence between groups of organisms.

In clones or in bisexual organisms each individual or, respectively, each pair of individuals of a population can be the starting point of a new ancestor-descendant-lineage diverging from the stem population (Fig. 20). The rare occurrence in nature of a polytomous genetically divergent evolution of clonal organisms (simultaneously giving rise to several lineages) is due to the limited availability of resources (food, humidity, hiding places, space for descendants, etc.) that offer a chance for living for only a limited number of varieties of organisms. Therefore the systematist can use the genetic similarity as a criterion for the discrimination of groups of organism, the criterion of divergence is applicable. In the case of large genetic distances to other populations it is not difficult to identify the members of a clone, and it is justified to name these groups (**agamospecies**; examples: Fig. 17). In case of smaller distances, however, disagreement may arise, and the naming of clones therefore requires a convention. The latter does only exist for bacteria:

The discrimination of taxa of bacteria is based on the classification of cultivated clones. For a microbiologist the individual "species" is a group of individual organisms, which differs from others so markedly in its properties that it is required to give the group a proper name. It is desired that a

16rRNA-sequence is published along with the description of the physiological and morphological characteristics of a new species. A rule of thumb is that a sequence difference of at least 1,5 to 2 percent is sufficient to establish a new species. Other values of similarity are given when DNA-DNA-hybridizations are compared (at least 70 % within species, less than about 70 % between species; this value is based on empirical experience). There exist exceptions (e.g., *Escherichia coli/Shigella dysenteriae*) where species are discerned that have a high genetic similarity but a markedly different physiology. Organisms that could not be cultivated and for which only the rDNA is known are classified as *Candidatus* (a provisional category) without species rank. About 99 % of bacteria in samples taken from nature have not been cultivated. The taxonomy of bacteria has mainly been designed to make the naming and handling of organisms in the laboratory easier (Brock et al. 1994).

The concept for the **biological species** (in contrast to the species concept in logics) presented herein in form of the **phylogenetic species concept** has the advantage to be testable because it refers to real historical processes and it is valid for clones as well as for sexual reproductive communities. Due to the properties of systems of the type "reproductive community" the point in time of the splitting of populations (see "transitional field between species", ch. 2.4) cannot be determined exactly. Cross-breeding experiments with organisms of recent populations allow the experimental test to find out whether the splitting is already irreversible or not. Furthermore, to recognize a species it does not matter at what time and in what number genetic novelties evolved within a population. The novelties could originate continuously and in any number. It is only important to know whether a divergence of populations occurred or not. It is not convenient to propose variations that are new to science (populations with unique characters) but that are not reproductively isolated as new phylogenetic species, because in practice it would not be possible to find universal criteria for boundaries between such species. In this case the question would have to be discussed what number of homozygous mutations occurring in a populations would be necessary to erect for a new variation a species. There exists no objective criterion for such a convention.

The phylogenetic species is the only category in systematics whose boundaries can be recognized without taxon-specific conventions in a phylogenetic tree. As already explained, it is not a system "species" that is a material entity of nature but rather the processes are real that cause the genetic divergence of a population from other such groups.

The phylogenetic species concept, which is a basic tool of phylogenetic systematics in the sense of Hennig (1950), is not in conflict with the usage of the species concept in weekday life. Therefore it is justified to use the term in the sense of colloquial language. For example, when we talk about a recent species, we mean organisms originating in a stem population and living today, independently of whether at the moment they form a functional reproductive community or not. We do not know the future and cannot decide whether an isolated population will be the starting point for the diverging evolution leading to a reproductively isolated population (= new species) or not. Therefore this possibility is ignored when groups are named. The idiomatic expression "species X is extinct", undoubtedly is wrong from ontological point of view, because a concept (a branch of a phylogenetic tree) is neither living nor dead. The statement is rather a short form for "the last members of the population that descended from the stem population X died." Misunderstandings, however, do not originate from this, apart from a misinterpretation of the ontology of the species concept. The phrase is useful due to its shortness.

Finally, to illustrate the problem concerning the ontological status of a species the question "**does a species 'horse' exist in nature?**" shall be raised. If we agree upon the **properties** an animal has to have to be called "horse", we can affirm this question: there exist many real organisms showing these properties and we can collectively name them in the sense of a logical class. However, is the horse as a group (a specific animal species) an "**entity of nature**"? The recent herds of horses are distributed over the whole world and do not form a functional reproductive community (although potentially all animals can interbreed). Only the single more or less isolated herd is a **real system** ("functional reproductive community", see ch. 2.2). Viewed in the four-dimensional space-time-frame, herds can be traced back to a stem-population and are the preliminary end products of a historical process. However, it is not possible to find objectively a distinct *natural* boundary representing the start of the process (start of the "evolution of the species horse", Fig. 24): the pop-

ulations of the different extinct organisms of the stem lineage of the recent populations of the genus *Equus* do merge smoothly into one another. If we **draw an artificial horizontal line** in this continuum and agree upon the convention that only those animals with specific properties clearly distinguishable in the fossil documentation are to be called *Equus* sp. (e.g., straight teeth with high crowns, complex enamel folds, long metarcarpalia/metatarsalia III, well developed intermediate tubercle at the distal humerus: MacFadden 1992, Fig. 24), we get a convenient convention. It serves above all communication between scientists on subjects such as

- the peculiarities of organisms,
- the period of time during which organisms with these properties existed,
- the point of time of divergence events, which could be evidence for changing environments,
- the affiliation of organisms to a material system.

Biological species: predicator for all node-free edges of trees that represent the complete real phylogeny. The delimitation of species in time requires conventions. The equation of this term with the *phylogenetic species* is the only possibility to classify the result of evolutionary processes of nature in an objective and universally valid way (see also preceding table with species definitions).

Phylogenetic species: a part of a phylogenetic tree between two irreversible divergence events, or a terminal species. A phylogenetic species is not an "entity of nature" but a mental concept.

Terminal species: a species to which recent populations belong to, or extinct species that did not produce daughter species.

Divergence event: process of simultaneous evolution of two or more irreversibly diverging populations originating from the same single ancestor organism or from a functional reproductive community.

Speciation: another word for an irreversible divergence event.

Finally some types of statements frequently used by biologists should be considered to realize that they are factually incorrect if taken literally (see Mahner & Bunge 1997): "I have studied the species *Anthura gracilis* " means "I have studied specimen which are classified as members of the species *Anthura gracilis*". The species itself is only a logical class; all members or "representatives" (not "parts") of this class, especially the deceased ones of past millenniums (which could well have looked somewhat differently) have certainly not been studied, but only single specimen of recent populations. – **Do species have properties**? Only an individual material thing can have a property. The statement "the species *Acrocephalus arundinaceus* (great red warbler) is markedly larger than *Acrocephalus scirpaceus* (reed warbler)" is wrongly phrased, because there exist, for example, small juveniles of the large species. Everybody naturally understands that such data on body size refer to the average or maximum size of adult individuals. The shorter expression serves the economy of communication. – "This species has typical yellow stripes on the back" means that we found yellow stripes on the individuals we studied and that we assume that all other members of this species show them as well.

Do species have signs of life? Species do not feed and mate, but only single organisms do. – The statement "species X evolved to a blind cave form" is not correct: since we agree that **species** are represented by branches between nodes of a phylogenetic tree and are composed of a large number of generations, species as an integral whole **do not evolve**. What is really changing during history is the presence of variations of genes within populations. In practice, however, every biologist understands what is meant with this statement. – The formulation "the mutation M originated in species X " is also comprehensible to everyone, although it is not exact: what is meant is that in the course of the development of a specific chain of ancestor-descendant-populations, a mutation first appeared in a single organism. This mutation has been transmitted to later generations and finally was present in all members of a population at a given time. – **Can species die?** As only individual organisms live, only these individuals can die. The terms "dying off of species" or "extinction of species" are metaphors intuitively understood by laypersons and biologist: "at the end of a species" all individual organisms assigned to the species die without having produced living descendants. – **Do species have descendants?** Only individual organisms can produce offspring, and these descendants are again individual organisms and not species. When

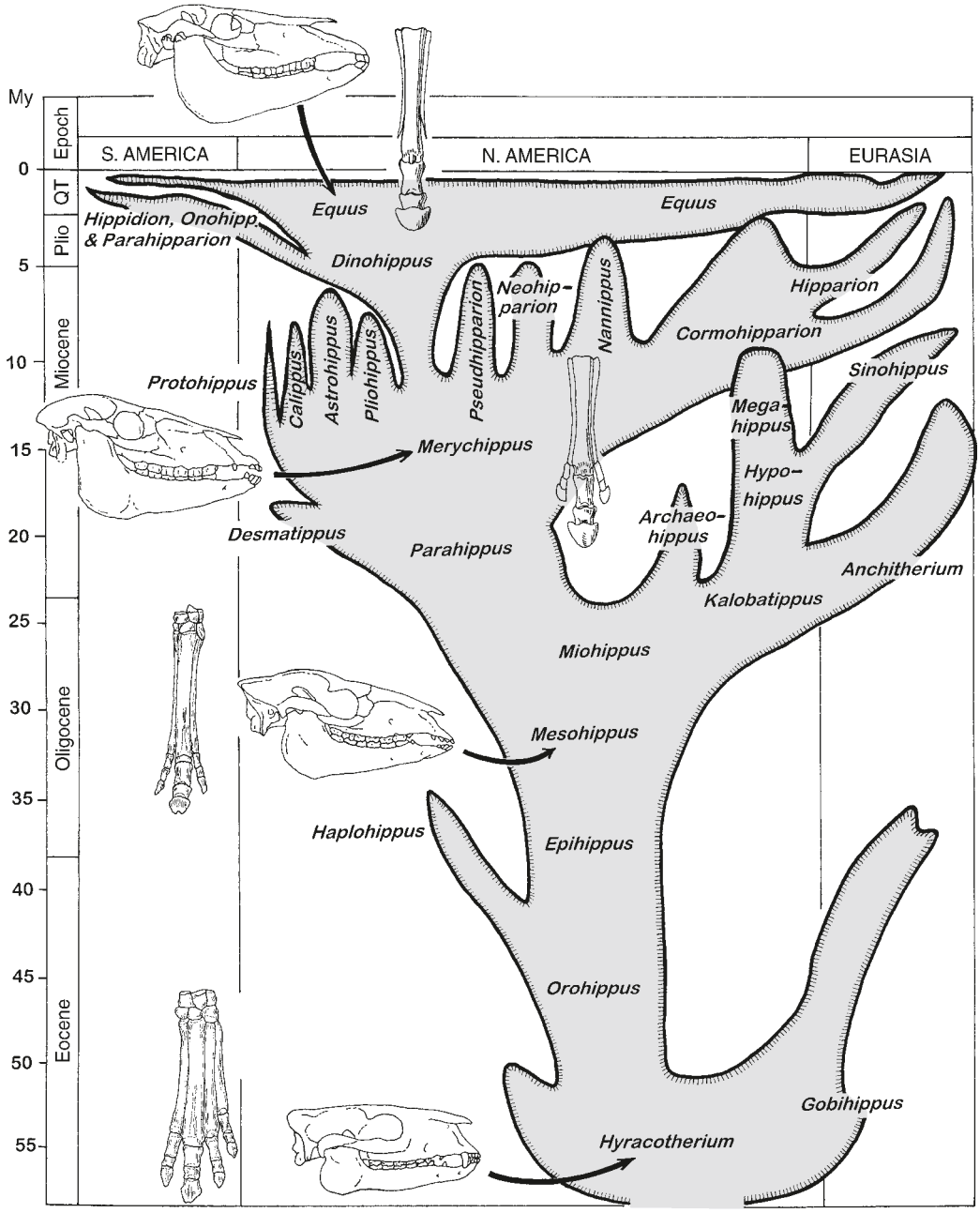


Fig. 24. Evolution of the Equidae (after MacFadden 1992). The series of ancestor-descendant-relationships is a continuum without natural boundaries.

"descendants of a species" are mentioned, it is meant figuratively that in a rooted phylogenetic tree a branch that represents a single species splits into further branches (daughter species). – When in the following chapters the "genes of species are compared", this phrasing is very simplistic and, strictly speaking, incorrect. However, no substitute has been used because the meaning is

easily understood by biologists. The assumption that a specific sequence is representative for a whole species is wrong in most cases. Usually sequences analysed in molecular phylogenetics were obtained from one or few individuals belonging to a recent population. Sequences may show variations in different individuals of the same population, and they can evolve within a species. Shortly after the final reproductive separation from the sister species probably all genes of the individuals of a population were more similar to those of the sister species or to those of the stem-population than today. Therefore, the genes present today are not those of "the species" in general. – The same inaccuracies occur in statements on taxa of higher rank. **Can a genus be discovered?** Someone saying he or she "discovered a genus" makes a wrong statement. Genera are man-made concepts which can be defined (ch. 3.5), and they also can be substantiated (ch. 2.6). However, the only thing that can be discovered is a yet unknown organism whose presumed position in the phylogenetic system does not fit within the genera so far defined. – The statement "the species *Helianthus annuus* and *Helianthus petiolaris* **can be crossed**, therefore the biological species concept is invalid" is a circular reasoning, because it presupposes that species may be known independently from the verification of reproductive isolation. This statement relies on a species concept different from the biological or phylogenetic one. The same holds true for statements on the frequency with which species hybridize in nature (compare Arnold 1997): the populations in question may not at all fulfil the requirements of a true phylogenetic species.

A stem lineage (or the "edge in a reconstructed phylogenetic tree") is always limited. Therefore it can be said that a species has a beginning and an end. In this context Ghiselin (1966) advocates the opinion that **species are individuals**. What "ends" or "starts" in nature has to be elucidated studying successions of generations (Fig. 20, 24, 28): at the beginning we find no birth of an organismic individual or some other sort of new start. All we can note is only a gradual, "borderless" change. The "boundaries" do only exist in our species *concept*.

2.3.2 Recognition of species

Presupposing we are going to use the phylogenetic species concept, in bisexual organisms the only available direct evidence for the affiliation to a certain species is the observation of successful reproduction. A cross-breeding experiment can show the extent of reproductive isolation. As this is rarely possible, in practice indirect evidence for the existence of reproductive barriers or for gene exchange within a population is used to assign individuals to a species. Circumstantial evidence is also used to group clonal organisms to species. Circumstantial observations can only substantiate hypotheses, they are not hard "proofs" and can lose their value when new evidence is found. Such indications are:

- Congruencies and differences in the characters of organisms. The distribution of characters allows the differentiation of groups when intermediate forms do not exist.
- The presence of very specialized structures (e.g., specific copulatory apparatuses functioning according to the key-keyhole principle) or forms of courtship and other behaviour allowing reproduction only with "appropriate" partners.
- Sympatry of morphologically different groups that do not hybridize or without intermediate forms. (Sympatry: living in the same geographic region, at the same place). If there occur at the same place rather similar but, due to the discontinuous distribution of characters clearly distinguishable groups, the most obvious hypothesis for the lack of intermediate forms is the existence of a reproductive barrier. In this case each group belongs to a different reproductive community.
- Infertility of hybrids.
- Position within a reconstructed dendrogram: if individuals so far assigned to the same species form a para- or polyphyletic group together with other related species, they may potentially belong to different species. A source of error that has to be considered: para- or polyphyly inferred from the analysis of genes can also be the result of horizontal gene transfer or may be obtained due to insufficient information content of the alignments used (an alignment is a table or matrix in which homologous sequences are written in rows in such a way that each homologous

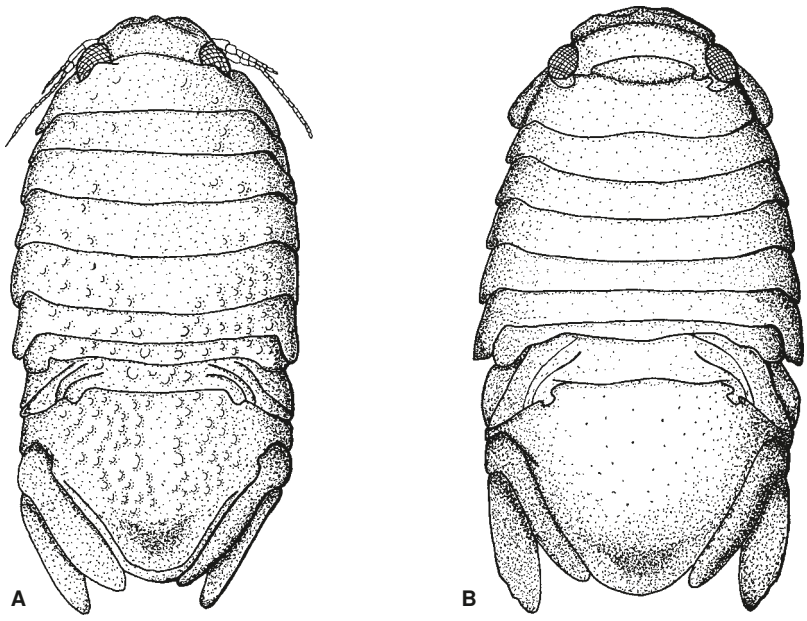


Fig. 25. Sympatric species of isopods (Crustacea, Isopoda) from brackish coastal waters of Northern Europe. The individuals can be assigned either to the species *Lekanesphaera hookeri* (A) or to *Lekanesphaera rugicauda* (B) due to the dorsal sculpture. Both can be found in the same locality. The absence of intermediate forms indicates that gene flow is interrupted.

position is arranged in a single column; see Fig. 103 and ch. 5.2.2.1).

– Genetic distance: it can be used in the same way as morphological characters. If two clusters can be recognized, each consisting of individuals more similar to each other than to those of the other cluster, and if intermediate forms are missing, then there is reason to assume that reproductive barriers exist or, in case of clonal individuals, that populations are ecologically separated. An absolute distance value suitable as criterion for the existence of a reproductive barrier cannot be given.

The taxonomist having only morphological characters at his or her disposal discovers patterns of character distributions which allow a grouping of organisms according to their visible properties. Congruencies within a group and the lack of intermediate forms between the members of different groups serve as evidence for the existence of separated functional reproductive communities (Fig. 25), which is equivalent to the identification of representatives of different species. This procedure is technically simple and effective. Such "**morphospecies**" however, have often been differentiated erroneously, when dealing with geographical races, different sexes, or with ontogenetic stages of the same species (Fig. 21). Experience is necessary to estimate the extent of intraspecific variability that can be expected for members of a taxon.

The interpretation of genetic distances to differentiate species is not convenient when only few individuals were analysed and when genetic differences are small, because some mutations may be typical only for single individuals. Furthermore, it is possible that two allopatric populations differ genetically, so that an analyses will show two clusters even when in case of a contact these organisms would produce fertile offspring. In plants this is relatively common, but this phenomenon also occurs in animals. Examples: the sequence of the COI-gene of populations of *Pollicipes elegans* (Crustacea: Cirripedia) from Southern California and from the Coast of Peru show 1.2 % differences due to reduced gene-flow. Individuals can be clearly referred to the local populations (Van Syoc 1994). – Many marine species of the West Atlantic coast of North America can be clearly separated into different local populations,

one group occurring in the Gulf of Mexico and the other along the east coast of Florida and further north.

When the genetic differentiation of different populations has progressed further, often "species" are prematurely named, although the genetic divergence is not irreversible because fertile hybrids do still originate. For example, the fruit flies *Dacus tyroni* and *Dacus neohumeralis* differ in colour and behaviour (individuals of *D. tyroni* mate in the evenings and those of *D. neohumeralis* during the day); nevertheless hybridization has been observed (Lewontin & Birch 1966). – In nature hybrids between blue whales and fin whales occur (Arnason & Gullberg 1993). – The Darwin finch species *Geospiza fortis* (medium ground finch) has a strong beak and can crack hard sees, whereas the smallest species of ground finches of the Galapagos Isles, *G. fuliginosa* eats soft seeds. Although these differences are maintained in nature, hybrids between these finch "species" occur. These "species" are not reproductively isolated (Grant 1993). If the status of a species is awarded to such populations, this implies the assumption that these populations will diverge further in future and will definitely separate into isolated lineages. Of course, such a prediction is unfounded. It could as well happen that due to changes in the rainfall regime on the Galápagos the ecological separation of populations will end and the two types of finches will again fuse into a uniform population.

The problems we often have when it seems to be difficult to decide if related populations should be regarded as two young species or as two races of the same species can principally not be solved and are typical for the "transitional field between species".

2.4 The transitional field between species

> "On the view that species are only strongly marked and permanent varieties, and that each species first existed as a variety, **we can see why it is that no line of demarcation can be drawn between species**, commonly supposed to have been produced by special acts of creation, and varieties which are acknowledged to have been produced by secondary laws." (Darwin 1859)

A reproductive community can split into two like a bushfire, the new systems can develop further into different directions (ch. 1.3.2). The moment of separation can be a very slow, potentially reversible process. In most cases it will not be possible to determine with precision from which moment onwards the divergence of populations is irreversible: the future of only potentially crossbreeding populations which are physically or ecologically separated cannot be known. In case of historical speciations it is in practice also impossible to determine the exact point of divergence. This is one of the causes for the problems biologists have when dealing with the differentiation of species. Whether the Java-mannikin *(Lonchura leucogastroides)* and the pointed-tailed mannikin *(Lonchura stricta)*, which currently produce hybrids (Clement et al. 1993), will diverge irreversibly or form again a uniform population in future cannot be guessed today. We have to accept that there exists a transitional field in which each effort to differentiate species "objectively" makes no sense; the endless discussions on this subject that can often been heard between taxonomists are fruitless. In this transitional field populations can occur forming "races", "race-circles" (Fig. 27), geographically separated "allospecies" (e.g., American bison/European bison) or evolutionary separated and not mixing "semi-species" even though they may have a geographical contact zone where they hybridize regularly, however without efficient introgression (European hooded crow/carrion crow).

Further examples: the South American fly species *Drosophila paulistorum* is divided into geographical populations which cannot be differentiated morphologically. Only some of them cannot cross breed with others: within this species some reproductive barriers are in *status nascendi* (Dobzhansky & Spassky 1959). Among the sittellas of Australia (Sittidae: *Daphoenositta*) five morphs can be distinguished, each with its own main center of distribution (Fig. 26): one could consider them separate species if they would not move out of their center of distribution and hy-

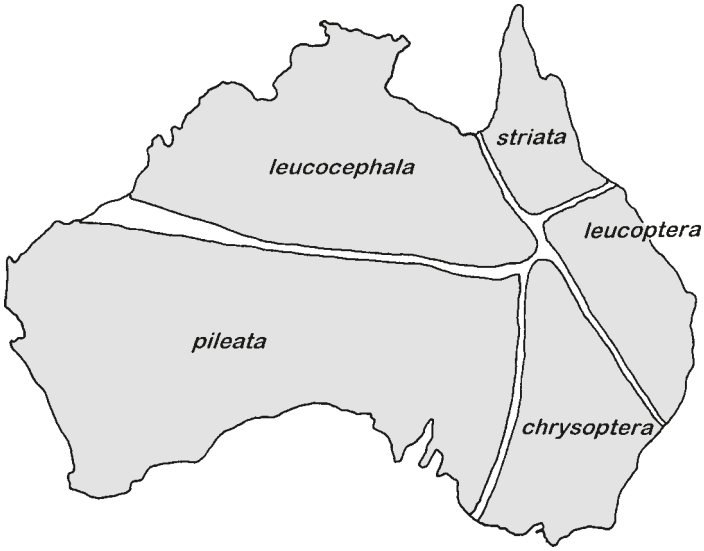


Fig. 26. Distribution of the Australian nuthatch species of the genus *Daphoenositta* (after Cracraft 1989).

bridize in Queensland. The local morphs undoubtedly are adapted to local conditions and only the complete interruption of gene flow between populations of the different morphs is needed to complete speciation. However, a future change in climate and subsequent growth of a more uniform vegetation in most parts of Australia could still cause the amalgamation of morphs. The point of controversy whether these morphs are races or starting species is fruitless because a glimpse into the future is impossible. The Australian nuthatches are in the transitional field.

Around the Swiss Jura dwell several populations of a millipede (Diplopoda: *Rhymogona montivaga*) that differ slightly from each other. They are partly understood to be races, partly regarded as singular species by different taxonomists. Presumably the populations were separated for a long time during the last Ice Age and were only able to return to the mountains after the glaciers retreated. A genetic analysis shows that there are five different populations. Each of them is most similar to the neighbouring population and together they show a circular gradient of genetic similarity. The ring closes in the Swiss Jura, where hybrids between the genetically most distant populations occur (Scholl & Pedroli-Christen 1996; Fig. 27). The hybrid zone shows that gene flow is still occurring. Therefore, it is not convenient to distinguish several species

Among iguanas of the Galápagos Islands sporadic hybridizations occur between marine and terrestrial iguanas (Rassmann et al. 1997). Both species are classified in separate genera *(Conolophus* and *Amblyrhynchus)*. As no introgressions are detectable, a future mixing of these animals, which differ in the way of living and in appearance is not to be expected; the populations probably belong to diverging species that are beyond the transitional field. However, as already mentioned, in the finches of the Galápagos Islands hybrids and introgression of genes have been observed (*Geospiza fortis* × *G. fuliginosa*, *Geospiza fortis* × *G. scandens*: Grant 1993), so that obviously the "species" could merge again under more uniform environmental conditions. The divergence of these Galápagos finches is reversible. Other cases that can be discussed are the hybridization of wolf and coyote or of bison and cow in North America.

Attention: A consequence of hybridization leading to fertile offspring can be the interruption of the divergence of populations (an inaccurate phrasing would be "boundaries between species are wiped out"). Another case is the origination through **allopolyploidy** of a third population that differs morphologically from the parent populations (allopolyploid individuals are (mostly tetraploid) hybrids or "addition bastards" of diploid parents, probably often originating from diploid

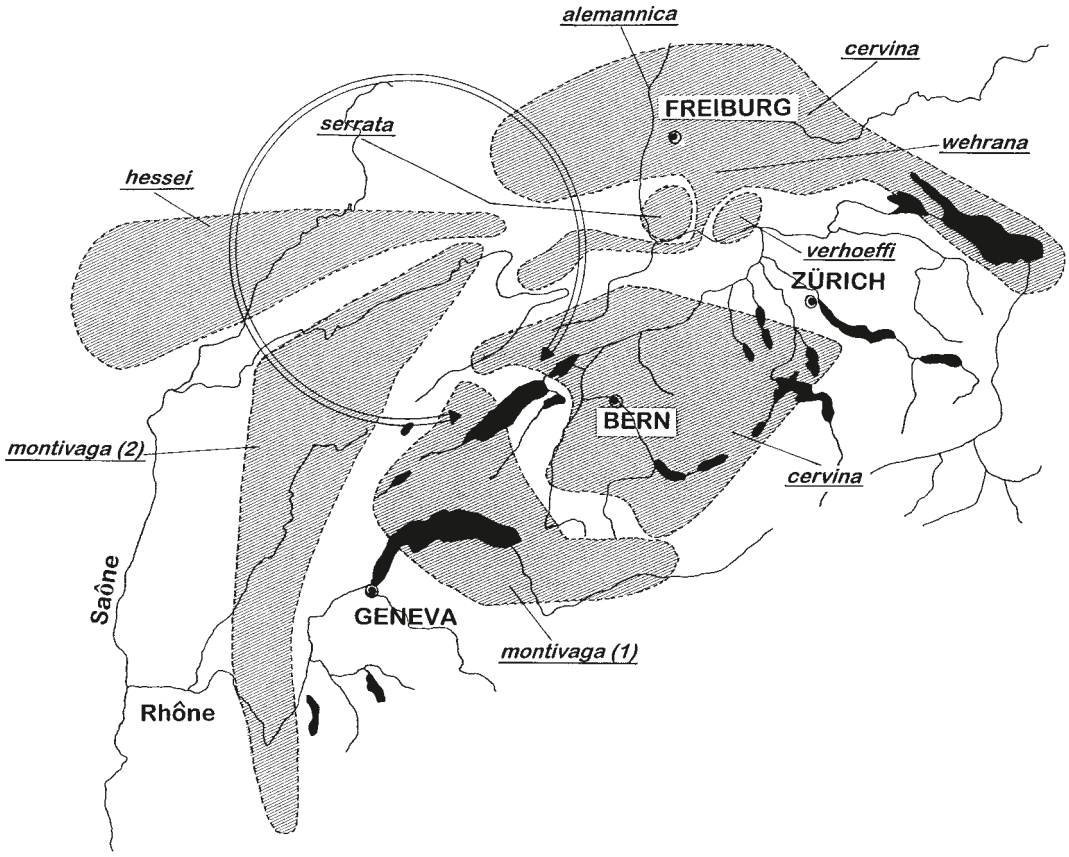


Fig. 27. Incompletely isolated populations ("races") of the diplopod *Rhymogona montivaga* form a ring closing in Switzerland (modified after Scholl & Pedroli-Christen 1996).

gametes; the parents usually cannot cross breed with normal gametes. Without polyploidization bastards are usually sterile because meiosis cannot proceed normally). A "speciation through hybridization" takes place should the hybrid population evolve further independently. It is also possible that hybrids originate regularly; in this case the gene flow with parent populations is not interrupted though it is unidirectional, and thus no independent evolution of the hybrids can occur (case of the European edible frog *Rana esculenta*, a hybrid of pool frog *Rana lessonae* and lake frog *Rana ridibunda*). In angiosperms the formation of hybrids through allopolyploidy is a very common process. Also in this case one should consider that as long as a network of gene flow through hybridization exists, due to methodological reasons (recognition of monophyla with the help of autapomorphies) a single phylogenetic species can only be defined for those stretches of the phylogenetic network in which ancestor-descendant lineages were isolated without introgression of genes from populations that are named differently. For other situations the phylogenetic species concept is not applicable.

For cases of hybridization as *Rana esculenta*, where a form of animals depends on the existence of two other species whose gametes are "stolen", the term *klepton** is used to characterize the status of this group in the system (see e.g., Crochet et al. 1995). These groups are not isolated evolutionary lineages.

* κλεπτης = greek for "thief".

2.5 Speciation as a "key event"

2.5.1 Notions and real processes

The term "speciation" turned up several times in the preceeding text. We have already seen that it is a process that may have different causes but always the same effect, namely the irreversible genetic divergence of populations. Note that we do not need a species concept to describe this process! Fig. 23B indicates that it plays a central role in phylogeny, because it causes the splitting of reproductive communities or the divergence of clones, phenomena that are visualized as ramifications in phylogenetic trees. The **genetic divergence** can be caused by several real processes (random accumulation of different mutations in populations separated in space and/or time; differing natural selection of properties in separated habitats (allopatric populations) or in different microhabitats at the same place (sympatric populations); sudden emergence of parthenogenetic populations through mutation or hybridization). In each case the result is an increase of the genetic distance between members of two diverging populations. The speciation is metaphorically also called **cladogenesis** ("origin of branches", actually the starting points of new species). The different mechanisms which may cause speciations are explained in textbooks on evolutionary biology.

For systematists the knowledge of the material objects involved in this process is relevant. Note that **evolution affects organisms and real systems, not concepts** (such as the species concept). The individual processes that are summarized with the term "evolution" affect the state of the system "reproductive community" and of clonal populations through the production of modified genes and by selection of individuals. The genetic divergence of populations, visualized objectively, for example, as increasing pairwise genetic distances between individuals, is the result of these processes: in bisexual organisms novelties lacking in sister populations propagate within a functional reproductive community. In innate "reproductively isolated" clonal organisms only selection will control the survival of a new life form or the stability of a "well-tried" and successful, i.e. well adapted one. ("Life form" stands for the average phenotype and the average way of living of the organisms of a population).

The consequences of the processes summarized with the term "speciation" are visible in an individual organism, the "carrier of the novelties". Therefore, it is possible to identify a single individual as a specimen of a new species. However, whether indeed the processes lead to the definite separation of sister populations or not, i.e. whether the speciation is completed depends on the development of the populations and not on a specific individual. If genes that enable mating between individuals of different populations are only present in few individuals, this could eventually lead to the fusion of two populations which already were composed of a majority of reproductively isolated members. Also the spatial extension of populations could have a decisive influence on whether the divergence proceeds or if contacts between populations occur that increase gene flow between populations. This has to be recalled when the question is debated whether "the units of speciation" are the organisms or the populations.

The analysis of genetic distances *does not require a species concept*, but is concerned with the consequences of those processes that irreversibly change the gene pool of populations. Therefore there is no circular reasoning when the species concept refers to theses processes.

2.5.2 Dichotomy and polytomy

A divergence of populations that develops further to a speciation produces at least two genetically distinguishable groups (Fig. 23A) each of which can be named a species. The most common case of divergence of two groups in nature is depicted as dichotomy in dendrograms. However, it cannot be ruled out that in a specific period of time in peripheral areas of a distribution area of a species single populations diverge simultaneously and independently from each other, so that several species originate at the same time. This case of "multiple speciation" is depicted with a polytomy (see Fig. 49).

Examples: In Jamaica occur several endemic species of land crabs of the genus *Sesarma* which have only one ancestor common. These species show different ways of living, they develop in brackish or fresh water, some even in bromeliads or in snail shells. The very fast radiation took place in the Pliocene. The sequence of early speciation events could not be resolved. Therefore, it cannot be ruled out that several evolving lineages diverged simultaneously (Schubart et al. 1998). – A similar process probably occurred in East African lakes. Cichlids of the genus *Tropheus* diversified quickly after the first colonization of Lake Tanganyika, probably six "evolutionary lineages" originated nearly simultaneously. Most of them colonize limited areas of the lake (Sturmbauer & Meyer 1992).

2.6 Monophyla

A group of organisms sharing an ancestor only common to them is called "monophyletic". Monophyla are not material things or systems of nature, but groups that we compile and differentiate mentally according to specific, scientifically substantiated rules. Therefore, monophyla are constructs with a relation to reality (natural classes). The term monophylum is only an instrument of systematics used to differentiate organisms that share a history only common to them.

This concept of monophyly contains a vagueness: who or what is considered to be the "ancestor"? We can take into consideration (a) a single organism that gave rise to a clone, (b) a reproductive community that existed at a specific time, (c) a biological species.

Clones originate unambiguously from a single stem-organism. Reproductive communities, on the other hand, stem from a series of successive generations. Following in thought the line of generations "backwards" into the past, the line will merge at some time with generations of another monophylum (the sister taxon). As there exists a continuum of generations starting with the first living cell it has to be decided through conventions where the boundaries have to be drawn to define single monophyla.

It can be seen in Fig. 28 that in reality there exists a vast number of groups of organisms all being monophyletic if the definition for monophyly given above is accepted. For the practice of systematics it would be fatal to want to name all of these groups respectively. Obviously it is necessary to differentiate between "**monophyletic groups**" which can also be groups of individuals within recent species, and **monophyletic taxa**. For the systematist it is important to discern species and groups of species and to name them as monophyletic taxa. Only some selected monophyla should be given proper names to avoid a confusing inflation of names.

Let us presuppose at this point that species are composed of generations of clones or of potential reproductive communities originating from each other, and that monophyletic taxa consist of species (see species concepts in ch. 2.3). Then the question has to be settled how to choose the boundaries of monophyla.

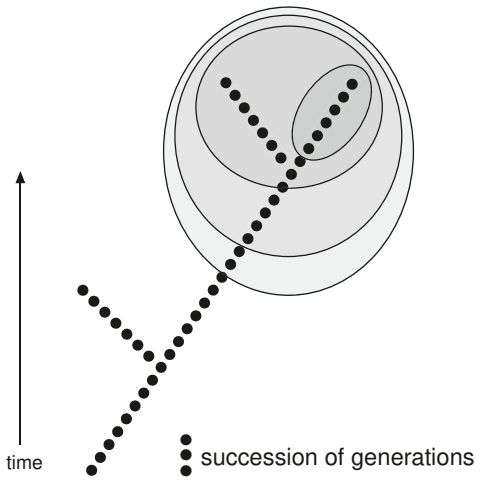


Fig. 28. How can a monophylum be defined in a series of successive generations? Each circle comprises a monophyletic group. There exists a vast number of such groups. It is impossible and not desirable to name all of them.

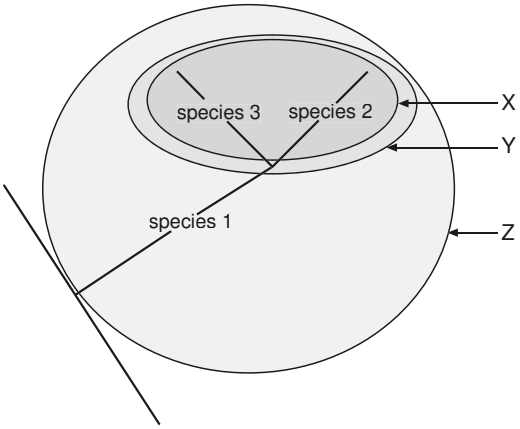


Fig. 29. Which is the correct demarcation of monophyla? The units X and Y comprise species 2 and 3, unit Z additionally species 1 (see text for explanation).

Monophylum: a monophylum consists (1) of a terminal species or (2) of a stem species and all descendants of this stem species. There are no descendants of the stem species that are placed outside the monophylum.

Monophyletic group: a stem-population (or a stem organism) and all its descendants.

Sister group (adelphotaxon, sister taxon): the closest related monophylum to a given monophylum in a dichotomous dendrogram.

Clade: (from greek χλάδος = branch, twig) branch of a dendrogram with all attached twigs and leaves (terminal taxa), independent of whether the topology is correct or not. A clade is not necessarily a monophylum (case of an incorrect topology).

Ground pattern characters: characters assumed to have been present in the last common stem-population of a monophylum.

In Fig. 29 theoretically four units of descent can be recognized when species (and not populations) are grouped:

Excluding single species from our considerations, the question remains whether the naming of the unit X, objectively comprising a different group of organisms than units Y and Z, would prove convenient in practice. If a group comprises unit X (Fig. 29), the common stem-population of species 3 and 2 is not included. If this stem-population is excluded, there could also exist other unknown descendants that are not included in the set X, whereby X would not be monophyletic. If the group comprises the set Y it certainly is monophyletic, but it already contains generations belonging to species 1! The groups with only species 3 and 2 can neither be demarcated clearly with set X nor with set Y. The next more inclusive monophylum going beyond the individual species and being unambiguously composed of species, comprises the sister species (2 and 3) and the common stem-species, or the **sister taxa** (= **adelphotaxa**) and the last common **stem-species,** respectively. The definition of the term "monophylum" (see below) results from these considerations. Since the first population that can be assigned to a specific species usually cannot be identified, we have to accept that a lack of resolution remains. The demarcation of monophyla cannot be more exact than the resolution within the transitional field between species (ch. 2.4).

The extent of a monophylum is determined with the naming of a stem species that existed in a specific period of time (ch. 4.4). The same group can also be identified when the sister taxon is named (Fig. 84). Within a species the demarcation of a monophyletic group of organisms requires the naming of a specific ancestral population or, in clones, of an individual ancestor. In bisexual organisms evidence has to be presented that a "monophyletic" population had no reproductive contacts with other populations since derivation from the indicated ancestors.

Because all members of a monophylum descend from a single last common ancestral population they share common characters they inherited from that population. Since organisms referable to a stem species usually are not known, in practice the analysis of the "last common stem population" and of the "stem species" has no importance, the characters of these ancestors have to be reconstructed indirectly. Not all ancestral characters are preserved during the course of evolution of the sequential populations. However, comparing the characters of all members of a monophylum, a set of characters which were already present in the stem-population can be reconstructed and distinguished from characters that originated later (ch. 6.2). Characters which are assumed to have been present in the last common stem-population of a monophylum are called **ground**

pattern* characters of the monophylum. A ground pattern is always a combination of several hypotheses (for further details see ch. 5.3.2). Example: most mammals have nipples and associated milk glands in the female sex. However, in monotremes nipples are absent. The reconstruction of mammal phylogeny shows that nipples probably do not belong to the ground pattern of mammals, but milk glands do.

"Real" monophyletic groups of species are groups which originated historically; they are natural kinds, but they are not material systems, because no processes take place between the parts of a monophylum ("the individual fires burn independently of each other"). The consideration of a monophylum as isolated "unit of nature" is a mental construct: the conceptional unit is a branch of the phylogenetic tree of organisms which has been arbitrarily cut at some place. Also the whole phylogenetic tree is a reconstruction, not an existing thing. The beginning of a monophyletic group is always defined with the period in time at which exactly the last common stem population (reproductive community, clone) or stem species existed (exception: two populations in the case of hybrids). Therefore, the resolution of boundaries can only be as precise as the identification of the period in time during which the last common stem population or stem species existed.

Attention: a "monophylum" can either be understood to be a monophyletic group of real organisms, or a mental construction of which we assume it represents a material group in the sense of the definition explained above. "Monophyla" in dendrograms are always hypotheses. Often parts of dendrograms are obviously artifacts of the method or of the data used and are not accepted by anyone as image of monophyletic groups. Parts of dendrograms can be termed "**clades**" to distinguish them from the unknown correct monophyla.

Note: sometimes other definitions of "monophyly" are used, for example in the sense of "sharing a last common ancestor", with no further specification. In this case a paraphyletic group is also "monophyletic". Whoever prefers this definition must use the term "holophyletic" in place of the concept of monophyly recommended in this book. Paraphyletic basal groups ("stem groups") could be named "orthophyletic".

Can a **species** be called "monophylum"? In the sense of the definition of the term monophylum only a terminal species (this is a species not ending with a speciation) can be a monophylum. By contrast, all internal sections between two points of speciation in a phylogenetic tree are not monophyla, because the descendants of these "internal species" would not be included in the taxon. This **paradox** is accounted for by the fact that the phylogenetic tree is a mental concept in a four-dimensional space. In reality, the "internal sections" have been series of generations which would have been classified as terminal species at the time of their existence before the next speciation happened. For methodological reasons, however, in the reconstruction of phylogeny we have to treat an "internal species" as part of a more inclusive monophylum.

In a philosophical sense, monophyla have properties of "**individuals**": they have a beginning and an end, they could be understood to be "historical entities". But what renders a group of species to be a "unit"? Obviously the organisms of taxa of a monophylum become extinct independently of each other, there exist no lawful reciprocal relationships between the members of the group. In our minds we combine carnivores to a monophylum, but there exist no interactions between a lion in Africa and a polar fox, material system relations are absent. Obviously, monophyla are units of our thinking, but they are not coherent objects or systems of nature. Only our intellectual concept (the term "Carnivora") has the property of an individual. In the same way a literary character is an individual.

How many monophyla are there? Taking the species as smallest unit: if there are N irreversible splittings of populations (speciations; ch. 2.3) in the phylogenetic tree of living organisms, then we get, including the terminal species

$$2N + 1 \text{ monophyletic taxa.}$$

This is much more than can be conveniently named. Practice decides which monophyla should

* the term "ground plan" is avoided because it suggests the existence of some planning authority; a ground pattern is always an incomplete and hypothetical reconstruction.

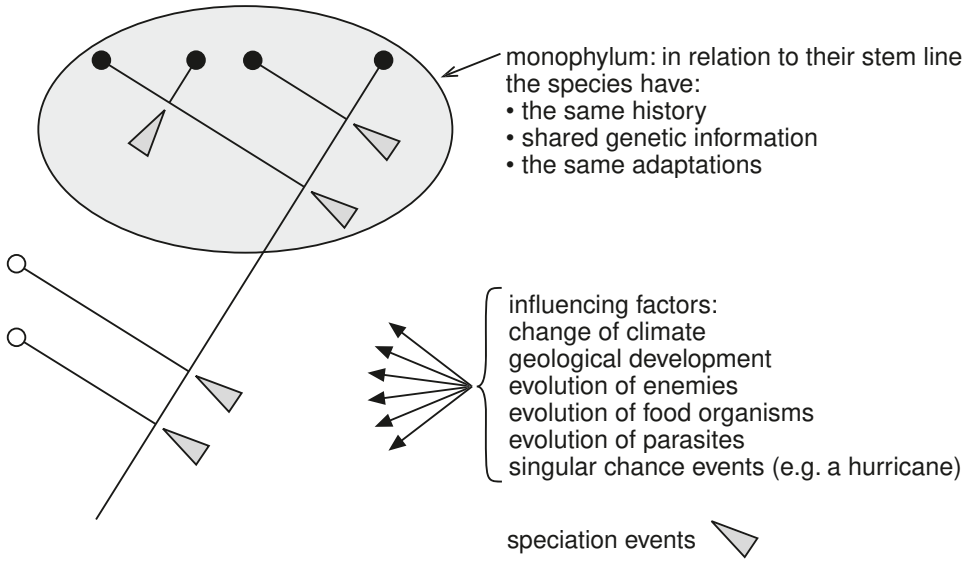


Fig. 30. Members of monophyla share common characters and the same history of descent.

get proper names to improve communication between scientists. (Especially ambitious systematists name as many taxa as possible, making communication more difficult).

Although monophyla are not material individuals, it is useful to recognize and differentiate groups of species of common descent. The **identification of monophyla** (ch. 4.4) has many **advantages** in biology:

- species assigned to a correctly identified monophylum share the same stem lineage as the monophylum and thus the same historical background which determined adaptations. Knowledge of the biology of one species of the monophylum often allows predictions about the biology of related species (Fig. 30).
- This means that members of a monophylum may carry common genetic information other organisms lack.
- Only if all the differentiated classes are monophyla, a classification of organisms is intersubjectively testable and an image of phylogenetic processes.

The definition of the term "monophyletic group" presented herein considers the practice to accept the existence of "monophyletic populations" within the limits of a species. It is not to be contested that there exist large populations that originate from a few founding individuals. Such a group can be called monophyletic independently of whether the origin of a new species within this group has been recognized or not, assuming that all descendants of the stem population had no reproductive contacts to neighbouring populations. In bisexual populations which are said to be monophyletic, however, it is difficult to prove unequivocally that single contacts (successful mating with individuals of neighbouring populations) did not occur. Furthermore this definition also allows to call groups of mitochondria and plastids monophyletic in order to do justice to the circumstance that organelles may evolve independently of their host organism. In addition, the proposed notion is independent of the species concept, which may be very different (ch. 2.3).

The identification and delimitation of monophyla will be discussed further in ch. 4.4.

2.7 Evolutionary theory and models of evolution as basis for systematics

When in the following sections the term 'evolution' is used, nothing else is meant but "change over the course of time ". Thus defined, the term is not linked to the adaptive value of a novelty. With this concept there is also "evolution" when neutral sequence parts change without having initially consequences for the phenotype. However, it cannot be ruled out that these "silent mutations" accumulate and sometimes sequences develop which have a new function.

Without the assumption that evolutionary processes occurred, a strong motivation for phylogenetic research is missing.

- Assuming the existence of evolutionary processes, the reasons for why phylogeny took place is obtained.
- The existence of recombination, of genetic drift, and of different selective advantages or disadvantages of novelties explains why only few novelties spread within populations.
- Studying the mechanisms that cause genetic divergence of isolated populations one finds the explanation for the phenomenon that populations may have a characteristic gene pool, with the effect that members of different populations can be distinguished.
- Identification of "reproductive barriers" allows evolutionary biologists to explain why populations diverge genetically and why the network of gene flow between bisexual populations can be disrupted forever. This is one of the basics of the phylogenetic species concept.
- Ideas on the evolution of sequences are the foundation for model-dependent methods used for tree inference.

Wiley (1975) mentions three axioms which can be the prerequisite for phylogenetic analyses: (a) evolution is a fact and occurs, (b) there is only *one* phylogeny and (c) characters are inherited. However, it can be shown that statements on relationships are possible without knowledge of these axioms, as demonstrated by tribes of primitive people.

Imagine a scientist not knowing evolutionary theory and pursuing the task to analyse the causes for the similarity of living organisms: to begin with he or she will answer the question whether the similarity (e.g., in morphology and in the life history of lions and tigers) is a product of chance or whether it is more probable that a common cause that produced the similarity has to be postulated. According to the argumentation presented in ch. 5.1, the assumption that the similarities of lions and tigers "originate from the same source" is the most probable one. The comparison of lions and tigers would enforce the conclusion that there must exist nearly identical "blueprints" and forms of information transmission. In the next step one could choose between some deity that created life forms according to some plan or a more profane alternative. Not knowing what the alternative is, one could nevertheless reconstruct a system of relationships. Starting with the consideration that similarities caused by the same process ("meaningful characters") have to be distinguished from chance similarities, algorithms for a cluster analysis with unique meaningful characters could be developed. To do this no knowledge on evolutionary mechanisms is necessary. Grouping of organisms with maximum-parsimony-methods (ch. 6.1.2) does not require assumptions on the processes that influenced the evolution of characters.

However, with this descriptive reconstruction of relationships no understanding of the driving forces is gained. Especially an **estimation of the plausibility** of the results (ch. 10) is not possible, which is especially necessary when the amount of available information is limited. Knowing how factors can drive evolutionary processes, it is understandable that haematophagous, ectoparasitic crustaceans living on fish could have originated gradually from marine carrion feeders (Fig. 180) and the plausibility of such a hypothesis can be explained. In contrast, evolution of blood-sucking parasites starting from specialized phytoplankton-filtering herbivores requires more intermediate steps in the form of further types of life forms. Such a hypothesis is less plausible, if

such intermediate forms are not known. In order to understand this, a scientist needs to have some expertise on evolutionary mechanisms.

Whereas an analysis of morphological similarities can do without assumptions on evolutionary processes, these are needed when sequences of molecules are analysed with distance methods and "maximum likelihood" methods. These methods require assumptions on evolutionary processes, which are integrated in the calculations of optimal topologies in form of models of sequence evolution. However, so far little attention has been paid to the question if in principle **evolution can be simulated with a simple model** and if models for the reconstruction of phylogeny are universally useful. According to the theory of neutral evolution (ch. 2.7.2.2), at least for molecular characters it has to be expected that substitutions in gene regions that are not under selection pressure occur stochastically. Substitutions should accumulate stochastically and therefore should be environment-independent whenever rate constancy in time and in different lineages (*rate stationarity*) is a prerequisite for the use of model-dependent methods of sequence analysis. In most cases this prerequisite is neither tested nor discussed, which can cause serious mistakes (see logics of deduction: ch. 1.4.2).

It certainly is true that the shape and physiology of organisms is to a high degree subject to selection and thus to changes of the environment. And it is indeed possible to describe the mutation- and selection pressure mathematically in order to predict the course of the development of a population when starting conditions (e.g., allele frequencies) and relevant environmental parameters are known. In practice, however, in most cases the variability and complexity of marginal environmental conditions that influence the evolution of populations cannot be recorded satisfactorily for recent species, and much less so for extinct species. Historical speciation events may have been influenced by very different and by a different number of environmental processes. These may include:

- changes in cosmic radiation,
- impacts of meteorites,
- volcanism,
- orogenesis (origin of new mountains that may create new climatic zones),
- drift of islands and continents (populations are separated),
- marine transgressions and regressions (landscapes are separated through marine ingressions, marine animals penetrate into karst regions etc.),
- climatic changes,
- extraordinary storms (they may carry for example insects to distant islands),
- appearance of new enemies, parasites, pathogens,
- occurrence of new food sources.

It is not predictable at which specific point in time of earth's history these events will happen. Even for a well monitored factor like the weather, it is not possible to get a reliable prediction for the temperature at Easter in Munich in exactly 20 years. It is possible to make statements on how probable it is that with increasing distance to the coast insects may reach an island (see MacArthur & Wilson 1967). However, it cannot be calculated which wind will carry at what time a pair of flies of a certain species to Hawaii. This type of prediction would be necessary to model episodic events in phylogeny. In reconstructing phylogeny we have to analyse real historical events, whereby in principle it has to be assumed that even the highly improbable (a pair of monkeys swims across a large sea on a single, free floating and fruiting tree) may occur. Since a consequence of these events often are radiations, extinctions, and changes in the morphology of organisms, we have to assume that the evolution of morphological characters is a **chaotic process** (as chaotic as changes in the weather), which takes a predictable course only for a short period and that can only be modelled when environmental conditions are surveyable for us (see size of bills of Galápagos-finches: Greenwood 1993). "Alter any event, ever so slightly and without apparent importance at the time, and evolution cascades into a radically different channel" (Gould 1989, p. 51).

Evolution considered over periods of times in which speciations occur, is not a mechanical process, for which the result would be estimated with a probability statement when the starting conditions are known. A living system receives, through the gain of novelties, unpredictable new properties which may change the rate of evolution. In this context Konrad Lorenz (1973) refers to the analogy of the totally new system properties of

an oscillating circuit, which cannot be deduced from the sum of properties of a capacitor and a coil. Because it is impossible for a scientist to get to know the limited, but in reality large number of historical causes which influenced the evolution of a lineage of reproductive communities, Lorenz concludes that the construction of higher systems (of living creatures) cannot be deduced analytically from the construction of lower ones. For the same reason the course of evolution of a character will not be inferable or predictable for long periods of time with the help of a model that is based on some evidence (e.g., rates in extant populations), when the character has a function and when it evolves non-neutrally. A stochastic course of evolution only dependent of mutation rates and genetic drift can only be expected for characters without function (e.g., pseudogenes), as long as the population does not experience an episodic catastrophic reduction of the number of individuals (this would also cause unexpected deviations of substitution rates). The assumption that a novelty is really without function would have to be tested.

By simulating changes of allele frequencies in populations it was possible to show that chaotic, non-stochastic evolution already occurs in virtual populations when marginal conditions are of little complexity. This happened even under constant selection conditions and more often with density dependent selection (Ferrière & Fox 1995). We must therefore conclude that evolutionary processes which cannot be described with precision using models will occur frequently in nature.

To sum up, it can be stated that it is not absolutely necessary to know the causes and mechanisms of evolutionary changes in order to reconstruct phylogeny. This knowledge is only essential when it is intended to estimate the probability that specific character transformations happen, or to describe a model for evolutionary changes. At present it is largely unknown which characters and taxa meet the requirements for modeling evolution. The phenomenological analysis (s. ch. 5.) used to reconstruct many essential aspects of the phylogenetic system of organisms during the 19th and 20th century manages without these axiomatic assumptions. However, in order to discuss the results of a phylogenetic analysis, especially the plausibility of a hypothesis of relationships, and to design a scenario of the evolution of a group of organisms, knowledge of the theory of evolution is indispensable.

2.7.1 Variability and evolution of morphological structures

The evolution of morphological structures is determined by

- variability caused by mutations,
- adaptations to varying conditions of life (climate, food, competitors, parasites, etc.) of all life stages through selection processes,
- the individual surroundings of organs in an organism limiting the number of possible modifications,
- the already available genetic information (the genetic make-up).

Selection pressure influences the variability and rate of evolution of organs: morphological structures are integrated in the whole "apparatus" of an organism, and therefore their individual variability is limited. Variations of an organ that do not fit in form or function are detrimental to the carrier of this character: the size of teeth has to be adapted to the size of the jaw; the diameter of the long bones of a leg has to be sufficient to carry the body's weight. Riedl (1975) talks of the **burden** of the present construction of an organism limiting the number of possible variations of an organ. This constraint acts as selection pressure which depends on the functional importance of a construction for the survival of the carrier. Not only the fit of an organ to the whole construction of an organism is a "burden", but also the adaptation to environmental parameters: for a bird gliding over long distances there are not many possibilities to vary wing construction without decreasing the efficiency of its air foils with a given air density and flight velocity. On the other hand, if a mutation does not cause a visible or measurable modification of the phenotype (the latter includes physiology!), we can expect that it will not be subject to strong selection pressure. Such mutations usually have no influence on the evolution of morphology, physiology, or behaviour, and can spread in a population unhindered, contrary to functionally important mutations which are mostly harmful.

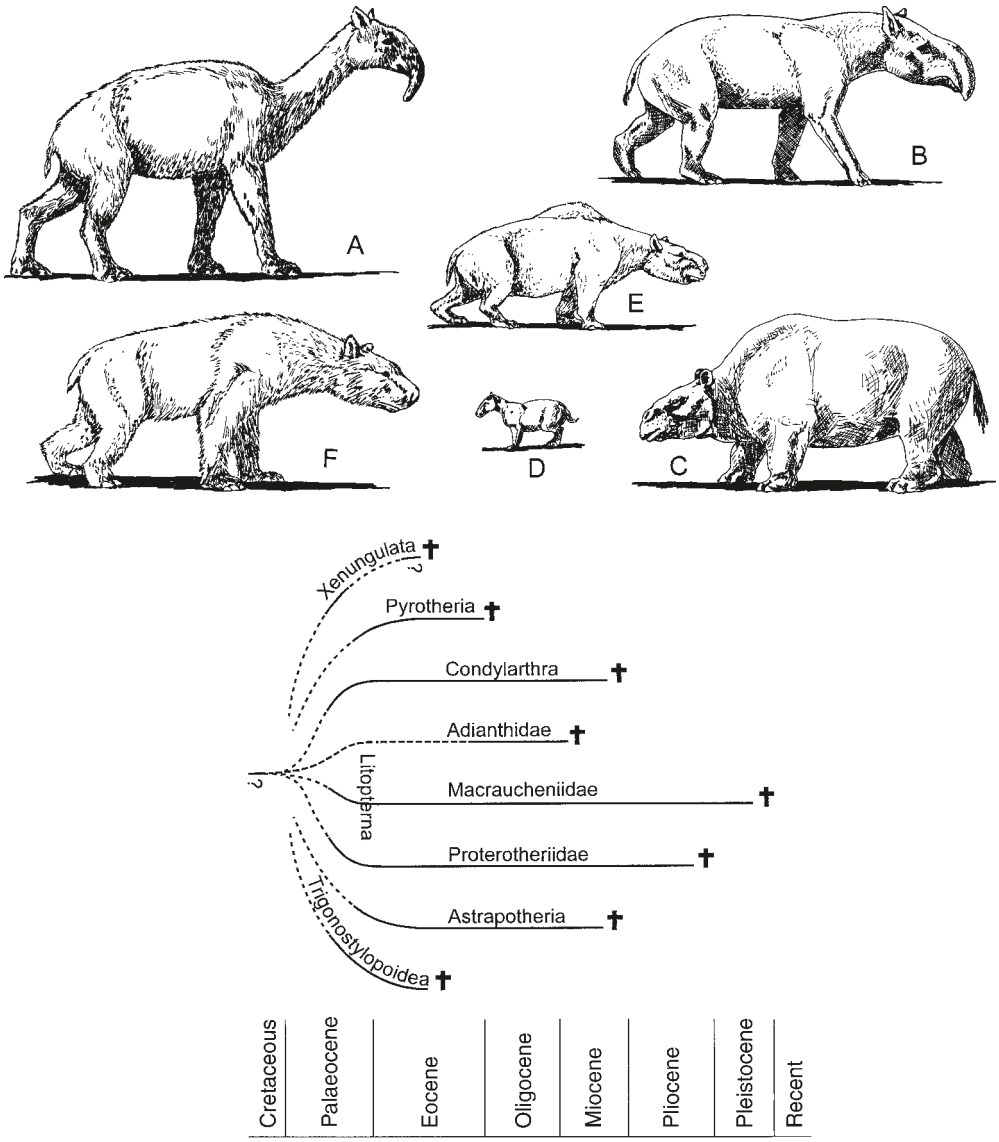


Fig. 31. After South America was colonized by ungulates that had the size of rabbits, a rapid evolution of new forms of life followed. They all subsequently vanished. Top: forms of South American ungulates (reconstructions). Bottom: radiation of ungulates (without Notungulata, according to Patterson & Pascual 1968). **A.** *Macrauchenia* (Pleistocene); **B.** *Astrapotherium* (Miocene); **C.** *Toxodon* (Pleistocene); **D.** *Paedotherium* (Pliocene); **E.** *Nesodon* (Miocene); **F.** *Scarrittia* (Oligocene).

It can be anticipated that temporary variations of selection pressure cause changes in the speed of evolution. The fact that the **evolutionary rate** of morphological characters varies in time is evident. Phases of very rapid adaptation to new environmental conditions and rapid appearance of new species alternate with long intervals of stable evolution. **The fossil record** proves, for example, that inconspicuous mammals coexisted with dinosaurs for millions of years, until an explosive development occurred in the Paleocene: ancestors of dogs, cats, monkeys and ungulates originated in a short period of time. Obviously, there is a connection with the extinction of dino-

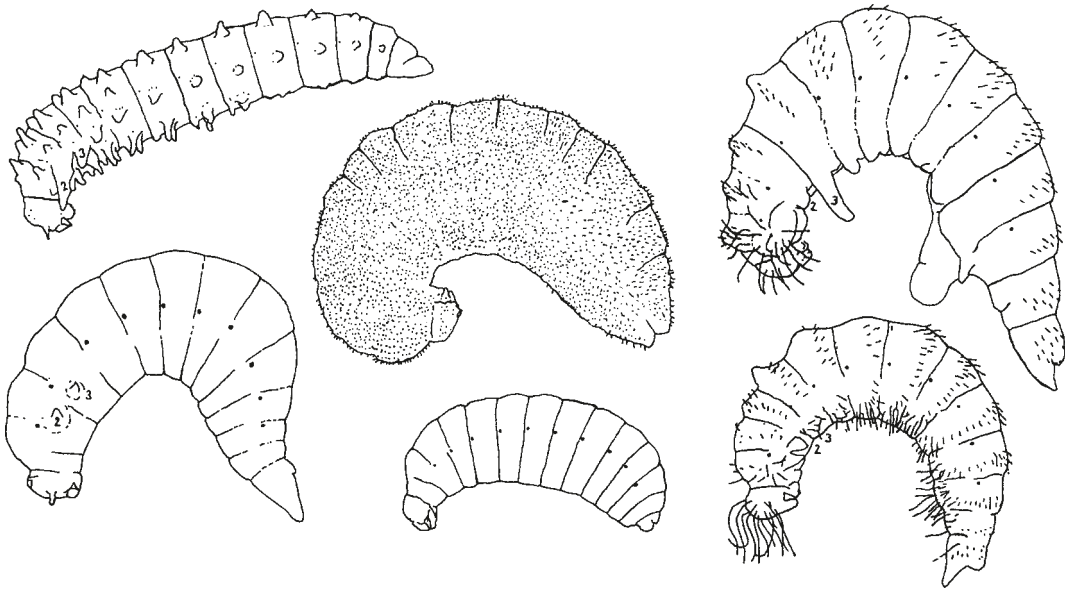


Fig. 32. Diversity of larval morphology of bees of the taxon Ceratini (after Michener 1977).

saurs: probably the resources that were not used any more by large reptiles could be claimed by mammals. However, it is not very probable that a biologist living in the Cretaceous would have been able to predict the extinction of dinosaurs and the future diversity of mammals, especially when the cause of the rapid change is the impact of a meteorite.

Besides fossils that are evidence for the irregularity of the evolutionary rate, further signs can also be found in the recent fauna, because the **different, taxon-specific variability of morphological structures** is very conspicuous. For example, most of the about 20,000 species of bees generally have similar larvae, maggot-like creatures growing up isolated in single honeycombs. The morphology of these larvae is subject to a stabilizing selection pressure, because they are probably optimally adapted to their way of life. However, in one monophyletic subgroup of Ceratini, no cell walls are constructed in the nest and several larvae live together interacting and competing for food. The effect of this different situation is that these bees show the greatest variability of larval morphology of all bee species (Fig. 32). In this monophylum the speed of evolution of larval morphology must have been substantially higher than that of adult morphology.

Single organs vary to different degrees: eyes of vertebrates, for example, are constructed in a very similar way from shark to humans, few variations are possible without hampering eye function. By contrast, within Bovidae there is a large variability of horns (bifurcated, twisted, long and short horns); birds of paradise have plenty of colours and exotic forms of fancy feathers on tail and head. The same region of the body is sometimes subject to very different selection pressures: within the cicadas the pronotum is a simple dorsal shield (as in nearly all insects); however, in the related tree hoppers (Membracidae) the pronotum has processes pointing backwards, or pointed elevations, or horn-shaped outgrows with globular or inflated parts; the biological function of these variations is not known (Fig. 33). The evolutionary rate of the pronotum shape is obviously much higher within the Membracidae than in other Cicadoidea.

Even though principally the mechanisms leading to an acceleration of the evolutionary rate are well understood, it will probably never be predictable whether and when for a specific group a radiation of species or a specific modification of organs will occur in a given geographic region. This is because the dynamics of all relevant environmental factors cannot be predicted, and in

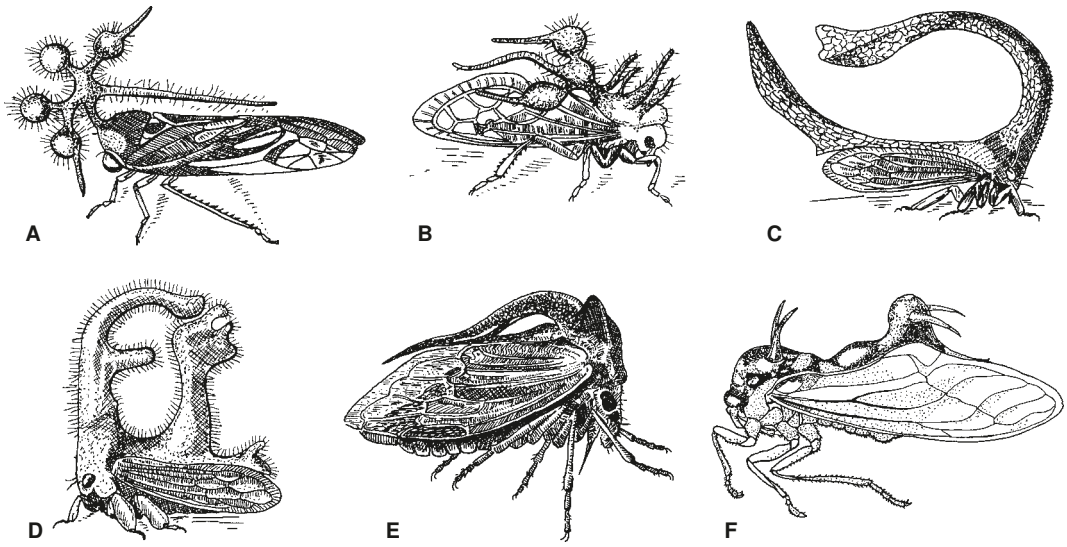


Fig. 33. Unusual variability of the pronotum in tree hoppers (Membracidae). Species of the genera *Bocydium* (**A**), *Cyphonia* (**B**), *Spongophorus* (**C, D**), *Centronotus* (**E**), *Heteronotus* (**F**).

many cases it even will be difficult to determine which factors are relevant.

A general phenomenon is the increase of structural complexity of organisms when a series of ancestors evolves rapidly. Species evolving slowly conserve archaic constructions and prove that there is (a) a difference in construction between conserved and more evolved organisms and that there are (b) differences in evolutionary rates. Recent coelacanth fishes *(Latimeria chalumnae)* have a body plan that already existed 350 million years ago, and the skeletal anatomy of frogs and salamanders has existed for at least 200 million years. In contrast, the diversity of ungulates originated in the last 65 million years, starting with small, somewhat hare-sized animals.

The evolution of complex novelties requires time: this is true for the development of technical instruments as well as for the evolution of organisms and their organs. It is not true in cases where "construction programs" are already coded in the genome and activated in a new part of the body. The starting point for anagenesis ("development to perfection") are organs and organisms with a simple construction: multicellular organisms originated from protists, Metazoa with mesodermal organs originated from organisms consisting only of external epithelia, Articulata with repeated segments evolved from unsegmented predecessors, arthropods with appendages in the form of specialized tools (claws, tweezers, mandibles, paddles etc.) are derived from arthropods with unspecialised appendages. The genetic foundations of evolutionary processes leading to the specialization of body plans are gradually being uncovered by developmental biologists. It is now known, for example, that the specialization of segments, of metameric organs and of legs of vertebrates and arthropods is controlled by several homeotic genes. They are similar to each other and obviously originated from gene duplication and stepwise evolutionary differentiation during the course of the phylogeny of the Metazoa (see e.g., Hall 1992, Shubin et al. 1997, Carroll et al. 2001). The morphological differentiation is reflected in an equivalent differentiation of genes (Fig. 34). To get an insect with three pairs of thoracic legs but without abdominal walking appendages, starting from a uniformly segmented milliped-like ancestor, obviously, amongst others, the development of thoracopods in the posterior part of the body has to be suppressed. One of the controlling genes for leg development is *Distal-less (dll)*. Its activity is suppressed in the abdomen of insects by products of the genes of the *Bithorax*-complex. But not all apparent reductions can be explained with the switching off of genes. For example, the evolutionary transfor-

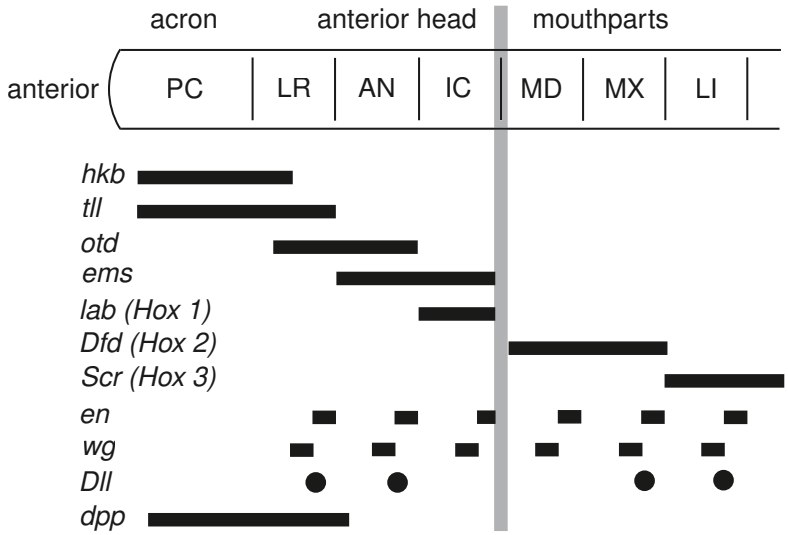


Fig. 34. In insect embryos a head is formed, with the sections acron (contains the protocerebrum *PC*), the region of the labrum (*LR*), which is not a true segment, the antennal segment (*AN*, contains the deuterocerebrum), and the intercalary segment (IC, contains the tritocerebrum) constituting the anterior part, followed by the three segments of the mouthparts (MD = mandible, MX = maxilla, LI = labium or maxilla 2). The differentiation of these segments is influenced by controlling genes, which are expressed regionally (indicated in the scheme by horizontal bars for the case of *Drosophila*). The engrailed gene (en) marks the posterior borders of the segments (modified after Cohen & Jürgens 1991).

mation of the hindwing of Pterygota to halteres in the Diptera is probably a consequence of many mutations of several genes that cooperate in a network, the expression of which are controlled by the *Ubx*-gene *(Ultrabithorax)*. A mutation of the *Ubx*-gene does not lead to an inversion of evolution, which would mean a back mutation of halteres to wings, but the consequence is the complete loss of the regulation of the development of halteres. Therefore, the phylogenetically older genes for wing construction are expressed and wings develop in place of halteres (Carroll 1994). The evolutionary formation of halteres has a much higher level of complexity than the apparent reversion of evolution through a simple mutation of the *Ubx*-gene.

Loss mutations, with which complex structures degenerate or are lost, probably originate often with only a few mutations. The "costs" for this change are therefore much lower than those for the evolutionary construction of these structures. This consideration is confirmed by the observation that damages caused by mutations (albinism, eye defects, deformed bones) occur more often in a given population than fundamental improvements of the performance of organs (sensitivity of olfactory organs, perception of ultrasound, digestion of cellulose, etc.).

As long as genes and embryonic developmental processes controlling morphogenesis are not known, it is not possible to make precise statements on the **genetic complexity of morphological novelties**. However, it can be assumed that the construction of a complex organ (e.g., of a statocyst) requires more genetic information than for the case of a simpler structure (e.g., a synapse).

Evolution is irreversible: from the point of view of physics, evolution has to be irreversible, because the evolutionary processes increase the entropy of the environment and this effect is not reversible. What the systematists call a "back-mutation" is physically not a return to the older state, but the origin of a new state which superficially is similar to the old one. What cladists use to call a "reversal" is either a mistake in the reconstruction of phylogenetic relationships, or a consequence of the inability to distinguish old states from seemingly similar new ones (see the above-

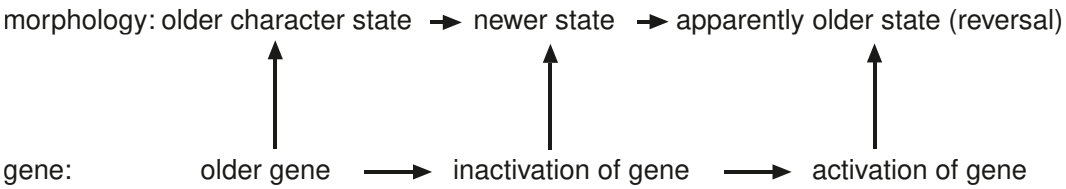


Fig. 35. Reversals of morphological structures do not have to be caused by back mutations.

mentioned example of the reversal of halteres to wings). As past conditions of life of an organism are the historical states of the environment, in the course of time the effects of historical environmental changes are irreversibly added.

"**Dollo's rule**" states that the evolution of organisms is irreversible. The irreversibility can often be observed: the fins of whales have a different structure than those of bony fish, the appearance of fins in a mammal is not a reversion to an older character state. Not a single aquatic mammal has secondarily acquired gills. As noted by Dollo (1893), a complex, phylogenetically older character state cannot evolve *de novo* once more from a younger one. In a weaker variant this rule means that the probability for the occurrence of a complex back mutation is lower than for the unique appearance of an evolutionary novelty. Therefore, it should not be possible that, for example, a blind deep sea crab that does not possess any more intact genes for the construction of eyes is the ancestor of a beach crab that secondarily shows (convergently) the same functioning complex eyes as the distant ancestors of the blind deep sea species.

Due to mutations causing losses, a character state can originate which apparently corresponds to a phylogenetically older state (e.g., loss of the cilium in unicellular eukaryotic organisms; reduction of the shell in conchiferous molluscs; loss of coiling of the shell in the snail genera *Fissurella*, *Ancylus*; in *Arenicola* or serpulids reduction of parapodia, which otherwise are typical for polychaetes). In cases where it is only a question of "switching off" genes, a gene will be detectable with techniques of molecular genetics even when it is not active, and the apparent reversal can be identified as a real novelty.

Proofs for the existence of inactivated genes have been known for some time: when a phylogenetically older structure, which was reduced and then has not been present for a long time in the evolutionary history of a taxon, suddenly appears again through abnormal mutations, this **atavism** indicates the activation of genes through mechanisms of gene regulation or back mutations. In these cases the expression of genes had been suppressed for several generations. Examples are the occurrence of three hooves on legs of recent horses, or the development of a short tail or of additional milk glands along the mammary ridge in humans.

The argument that evolution is generally irreversible is apparently not universally valid: a point mutation in DNA molecules, for example, can restore a phylogenetically older state of a single sequence position. This process will be noticed in the form of "noise" during the analysis of DNA sequences, because it can produce shared states (analogies) with outgroup taxa (see ch. 6.3.2). Comparing the molecule as a whole, however, it will be obvious that the phylogenetically younger molecule is different from the older one. Only a limited effect can be attributed to these **erratic back mutations**. They occur at random (in most cases probably not due to selection favouring a plesiomorphic state) and can only affect a small portion of the variability. Considering the surrounding sequence, it can easily be noted that the sequence has become different after some time, even when single back mutations occur. When many evolutionary novelties are present in a character, it is highly unlikely that a state identical with an older one develops anew through random mutations.

The analysis of the variability of bills of Darwin's finches shows that the evolution of morphological structures is often only apparently reversible: observations on the Galápagos Islands since the drought in the year 1977, documented a tendency of increasing body and bill size in the species

Geospiza fortis, which was probably propelled by food availability and sexual selection. When in 1983 the irregular climatic "El Niño" phenomenon started, the Galápagos Islands received plenty of rain and a vegetation developed producing small seeds. In this situation birds were favoured that had smaller beaks, and due to the reduced nutrient contents also a small body size was advantageous: the trend caused by dry years reversed (summarized in Weiner 1994). This case, however, is only an apparent reversal, because the effect is based on a shift of gene frequencies and not on the evolution of genes. If all genes for small bills were completely eliminated from the population after a drought period, it would take much longer to evolve smaller bills and the population would not be able to react to fast climate changes. At least at the level of genes after a phase of adaptive evolution, the state (of the genes) would not be the same as before.

This example also shows that morphological changes are only predictable, when a) it is known which adaptations to a changing environment are in principle possible in a species and b) when the future history of environmental factors is predictable. Only when these marginal parameters are known, it is possible to design meaningful models for the evolution of morphological characters.

2.7.2 Variability and evolution of molecules

The study of molecular evolution is relevant, because many methods used to reconstruct phylogeny on the basis of sequence information make assumptions about evolutionary processes (see ch. 8). As the structure of organic molecules is directly or indirectly coded by the structure of nucleic acids, for the analysis of molecular evolutionary processes it is especially rewarding to compare RNA- or DNA sequences, because the largest possible number of substitutions can be read only in nucleic acids. The degeneration of the genetic code, variations in codon usage and the existence of large non-coding regions have the effect that part of the substitutions that occur in nucleic acids cannot be identified any more in the proteome.

Since some sequence regions or single sequence positions often are subject to only weak or even

Fig. 36. Variations in the shape of bills in medium Galápagos ground finches *Geospiza fortis*. During periods of droughts, when only large, hard seeds are available, the average bill size increases within a population (after Weiner 1994).

to no selection pressure, random mutations have a greater influence than in morphological characters. Evolutionary changes of nucleic acids are the consequence of mutations, which are either neutral for selection processes or are exposed to weak or strong selection. This difference is of great importance for weighting of characters according to the probability of events (in this case: probability that specific substitutions occur) and for the design of appropriate models for evolutionary processes. Obviously, in all organisms molecules or sequence regions can be found which are more strongly conserved than morphological characters and thus can be homologized in macroscopically very different organisms (e.g., homology of rDNA regions in unicellular organisms, plants, animals; homology of vitellogenins in nematods, insects, vertebrates). The other extreme are those sequences that evolve rapidly, varying even within species (e.g., the control region of mitochondrial DNA, introns, satellite DNA, pseudogenes). Knowledge of these differences is significant for the selection of suitable genes for sequencing projects at phylogenetic or population genetic level.

2.7.2.1 Changes in populations

The location of a gene within the chromosomes is the **locus**, variants of a gene at a specific locus occurring in chromosomes or different individuals of a population are **alleles**. The number of diploid organisms of a population of a species that carry a specific mutated allele in form of a single copy (heterozygous) or with two copies (paired, homozygous) changes in the course of time. Some alleles are lost from the population after some time, others become more frequent and spread and finally some are found in each individual. In this case we say that the allele is

"fixed" in the population (see Fig. 5). The fate of individual alleles is determined by

– selection and
– random genetic drift.

Selection causes a **channelled** change of allele frequencies controlled by environmental parameters. This change is predictable when the population is large enough and when the effective environmental conditions and the contribution of alleles to the organisms' fitness are known. Each mutation in a gene that contributes to the survival and reproductive success of an organism can either improve, reduce or maintain the fitness of the organism in comparison with competitors of the same species. A mutation that does not influence or change fitness is called **neutral**. It has become a habit to talk about the fitness of an allele when the change of fitness of an organism achieved by the expression of this allele is meant. When a new allele is advantageous and dominant, after a few generations it will supersede the older allele in the population. From that moment on the novelty is **fixed**, it is a character of the whole population. We also say that a **substitution** has occurred.

When mutations are noted comparing individuals of the same species, the **mutation rate** (e.g., mutations per generation of a species) can be calculated for a gene in this species. For this purpose it is necessary to know the time span in which mutations are expected to occur. For small-scale studies it is interesting to know who was the common ancestor of mutated individuals in which the mutation occurred, and the period in which this ancestor lived. Such studies of mutations in humans are common medical research (see e.g., Gibbons 1998). **Substitution rates** can be inferred from these mutation rates in a precise way if exact analyses of population dynamics and of the selection pressure taking effect on specific mutants exist. The description of these highly complex processes is one of the tasks of population genetics. For the purpose of phylogenetic systematics such exact analyses are not available, because the time span that has to be considered and the number of species are too large. Therefore, substitution rates are estimated with statistical methods (see below and chapters 8.2, 14.1).

New mutations arise continuously in populations, and also repeatedly new alleles are fixed in a sequence of generations. However, new alleles are lost much more frequently (Kimura 1962). The probability that alleles are fixed depends on

– the original **frequency** of an allele,
– the contribution of an allele to the **fitness** of the organism,
– and the effective **population size**.

Whether and how fast a mutation that can become an evolutionary novelty disperses in a population depends, among others, on how large the difference in fitness of alleles is in a specific environment. When heterozygotes have the higher fitness value in comparison to homozygotes (heterosis, advantage of heterozygotes), the older allele is not displaced from the population, but an equilibrium of allele frequencies develops (e.g., sickle-cell anemia of humans). With codominant alleles, heterozygotes have an average fitness value of the corresponding homozygotes, homozygotes of one allele can have a higher fitness value than homozygotes with the other pair or heterozygotes. Furthermore, an allele being subject to no or to only little selection pressure (neutral or nearly neutral alleles) can behave like a strongly selected allele due to spatial coupling to another allele which is exposed to strong selection.

Allele frequencies can also change **at random** or without direction, because rarely are all alleles of one generation transmitted to the next one. These changes are called **genetic drift**. When all alleles are functionally equivalent and exposed to the same selection effects, allele frequency is determined only by genetic drift. Loss of alleles in a reproductive community may be caused by several factors: since each generation produces many more gametes than descendants, only a fragment of the alleles copied during gametogenesis will be retained. Furthermore, in diploid organisms part of the descendants carry homozygous genes and thus a smaller share of the available genetic diversity. Also, mortality fluctuates in populations with environmental factors. Catastrophic events can destroy a large portion of the original genetic diversity.

In smaller populations novelties are lost relatively fast because often rare new alleles are not transmitted to descendants. On the other hand novelties can also be fixed faster than in large populations, the speed of evolution is increased in this case (e.g., Li 1997, Ohta 1997). Catastrophic reductions of population size modify smaller

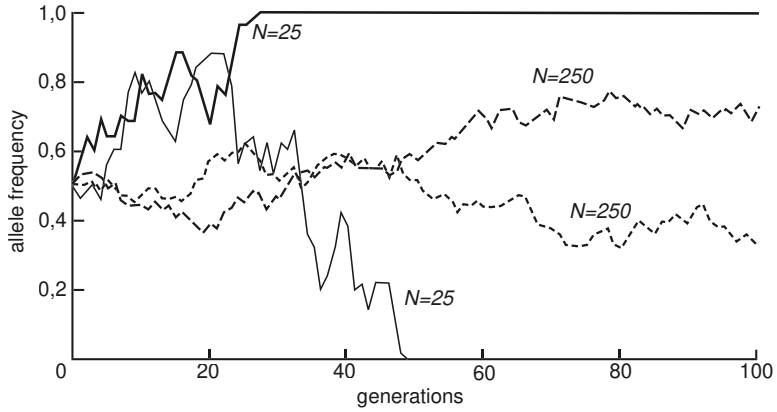


Fig. 37. Theoretical model for the change in allele frequencies due to genetic drift in populations of different size (N: effective population size; modified after Li 1997). Alleles are fixed when their frequency is 1.0, they are lost when the frequency is 0. The probability for the fixation or loss of an allele is much higher in a small population (N = 25).

populations more drastically than larger ones, more genetic diversity is lost. As the population size of a species can vary markedly during the course of its existence, the consequence can be a strong variation of genetic drift and thus also of substitution rates in short periods of time. When reductions of population size are caused by changes of selection parameters, those alleles that produce positively selected properties spread more rapidly. Such "bottle-neck situations" are important periods of fast evolutionary changes of a species. Effects of population bottle-necks have been demonstrated in several cases for phenotypic characters (e.g., Willmann 1995). Therefore, it has to be expected that **rates of molecular evolution** vary in a similar **unpredictable** way as for example the local weather that also influences the state of populations.

For **neutral alleles** in ideal populations the **probability of fixation** only depends on the drift rate. The substitution rate of neutral alleles in a population is identical with the mutation rate and independent of the population size (Kimura 1968). The last statement, however, does not hold for very small populations, which can be reduced to a few individuals by catastrophic events from time to time. The influence of the effective population size can be shown with a simple model that ignores drastic fluctuations of mortality and reproductive rate (Fig. 37). Genetic drift is much faster in small populations than in large ones. When the population size fluctuates, as usually happens in nature, the drift rate also changes. The substitution rate of alleles with **non-neutral mutations** depends on population size and in addition on the selective advantage or disadvantage of the mutant. These considerations are fundamental for the theory of neutral evolution (see below).

Systematists must take into account that for historical populations whose fluctuations are not observable, neither the drift rate nor the fitness value can be estimated for new alleles, because the effect of selective advantages and of population size cannot be ascertained. In populations whose size does not shrink occasionally to a few founder individuals, it is expected theoretically that neutral mutations spread with a uniform rate that is independent of fluctuations of population size. If an allele codes the optimal structure of a protein, in a large population the probability of fixation of a different, suboptimal new variant is extremely small. Therefore, sequences coding for well adapted phenotypical characters can be preserved over hundreds of millions of years.

2.7.2.2 The theory of neutral evolution

Up to the end of the fifties of the 20th century, evolution was understood as the result of mutation, recombination, migration, and above all selection. The evolutionary moulding of characters and the speed of evolution were thought to

be dependent above all on selection. This paradigm resulted from the analysis of morphological characters which are shaped by environmental factors. With the progress of molecular genetics new data accumulated which prove that molecules may evolve even without the influence of selecting agents. Polymorphisms were detected which are not visible in the phenotype. The theory of neutral evolution that was developed from single empirical observations and based on theoretical considerations is today an important element of evolutionary theory. The content of the theory should not be inferred from its name: on principle, evolution is not neutral, but it can be shown that there exist mutations that are neutral to selecting agents.

The theory of neutral evolution states that a large portion of changes occurring in molecules can be explained with random mutations which are not (or only to a small degree) subject to selection pressure. A more general statement is the well founded hypothesis that the substitution rate of a character depends, among others, on the selection pressure acting on a novelty. Therefore, mutations that are "neutral" will be retained more often in a population than non-neutral ones, because most non-neutral mutations have damaging effects.

This theory was developed for the evolution of molecules. However, the findings described above are also of fundamental importance for the understanding of the variability of morphological characters.

Substitutions which are neutral to selecting agents accumulate only in dependence of the mutation rate (Kimura 1968, 1983, 1987). If mutations hit the genome at random, the frequency of occurrence of neutral substitutions can be predicted when the mutation rate is known.

The following mutations are neutral:
- mutations that do not cause changes in the protein coded by a gene (**synonymous substitutions**, Fig. 48), mutations which do not influence translation or the function of RNA-molecules,
- mutations in amino acid sequences that do not modify the function or efficiency of a protein,
- mutations in sequence positions which have no function.

The assumption that neutral evolution exists does not imply that all alleles have the same fitness, but that for neutral mutants genetic drift has a larger influence on allele frequency than selection. Genetic drift also influences the dispersal of mutations that are subject to selection. A consequence of these mechanisms is a fairly **constant substitution rate** in periods of constant population size, because the largest part of mutations occurring in a population is neutral. It is assumed, for example, that in eukaryotic organisms the substitution rates of the proteins aldolase C and triosephosphate-isomerase (TPI) are relatively constant and, depending on the calculation, for aldolase is on average about 0.23 to $0.29 \cdot 10^{-9}$ substitutions per position and year and for TPI 0.30 to $0.42 \cdot 10^{-9}$ substitutions per position and year (Nikoh et al. 1997). However, one should not forget that over longer periods of time changes in cellular processes, such as modifications in the repair mechanisms for DNA or different concentrations of tRNAs, as well as differences in metabolic rates, generation times, and drastic fluctuations of population size may cause irregular substitution rates for neutral mutations.

A high portion of an organism's DNA is not transcribed (in mammals >90 %) and probably evolves neutrally. A large portion of mutations is probably not subject to selection. However, caution is necessary. Introns and pseudogenes are supposed to be without function, but this is not always true in the case of introns. The conservation of secondary structures in non-coding areas can be of importance for activities of proteins involved in gene regulation, for example. Furthermore, one has to keep in mind that functionless regions of DNA are not suitable for phylogenetic sequence analyses, because such regions become noisy rapidly (see chapters 1.3.5, 2.7.2).

Kimura's original concept of the theory of neutral evolution has been modified several times. According to Ohta (1973, 1992), most mutations that are not intensively selected are mildly damaging and not strictly neutral. Models of population genetics suggest that for this reason genetic drift is especially effective in small populations and the rate of evolution is higher (Ohta 1997), whereas in large populations selection is effec-

tive even against mildly disadvantageous mutations.

An estimation of the probability that characters are informative, which is very important for systematists, can be derived from this theory: considering taxa with long divergence times, one has to choose characters which are subject to stronger selection pressure, because these should evolve more slowly and thus retain with higher probability apomorphies that have not been substituted (we could also say that the phylogenetic signal would not erode so fast). These generally are characters that do not evolve neutrally, and this fact implies that substitution rates may depend to a high degree on the stability of environmental parameters. However, if taxa with short divergence times are to be analysed, we have to choose characters which presumably are neutral or nearly neutral and thus evolve rapidly. Evolutionary rates are discussed in ch. 2.7.2.4.

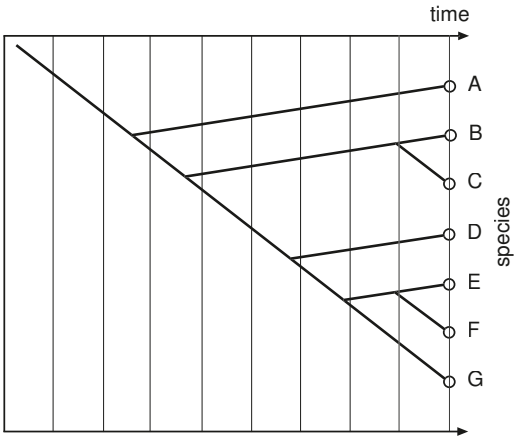


Fig. 38. Graph illustrating the effect of a molecular clock: spaces between vertical lines represent a unit of time and simultaneously the same absolute number of substitutions per unit of time. A reconstruction of phylogeny based on a precise estimation of the substitution rate would be very simple.

Neutral position: a sequence position which is not exposed to selection. An exchange of nucleotides or amino acids does influence the organism's fitness.

Synonymous or silent substitutions: a substitution in a coding DNA sequence which does not cause a change in the corresponding amino acid sequence. Synonymous substitutions do not have to be neutral all the time (see ch. 2.7.2.4).

2.7.2.3 The molecular clock

When sequences evolve neutrally, changes occur even over millions of years with a predictable, constant average rate. The neutral evolution of sequences is characterized by chance events and is independent of selection processes. The occurrence of neutral substitutions is comparable to the decay of radioactive elements: in short time intervals the frequency of individual events is not predictable with precision, but for longer periods predictions for the total number of events are relatively accurate and differences between parallel observations of the same process are small. It is interesting to assume that a molecular clock exists, because it would allow the reconstruction of phylogeny with simple distance methods (Fig. 38; distance methods: ch. 8.2).

Assuming that the occurrence of substitutions is a stochastic (regular and random) process, it is possible to describe sequence evolution with a model and to calculate divergence times and the age of monophyla from genetic distances (Figs. 38, 169). The true genetic distance d is the total number of substitutions which occurred in the lineages that separate the sequences of two species (also called the path between two terminal sequences). Assuming that the substitution rate λ and the estimated distance d between the species (or between the last common ancestors of two monophyla) are known, the divergence time t of two species (or of 2 monophyla) can be calculated with the formula $t = d/(2\lambda)$ (Fig. 38, 41, 163, distance analyses: ch. 8.2). However, attention has to be paid to the following observations:

– sequence positions evolving neutrally and with a constant substitution rate are also those that are modified rapidly and therefore get noisy faster than more conserved sequences due to multiple substitutions. This means that with high probability they conserve evolutionary novelties only for a short period of time (for the term "noise" see chapters 1.3.5 and 4.1). A typical indication for the occurrence of multiple substitutions is a low ratio of transitions to transversions in pairwise sequence comparison (Fig. 39; see also Figs. 42, 43).

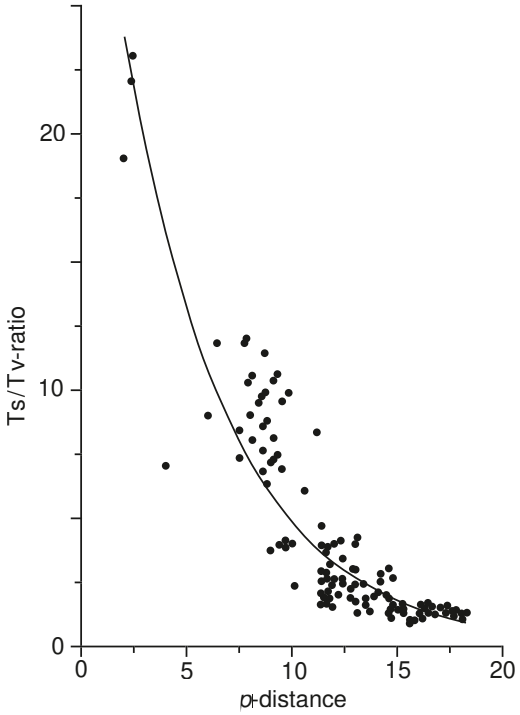


Fig. 39. Correlation between the ratio of transitions to transversions (Ts/Tv) and the uncorrected ***p***-distance visible in pairwise sequence comparison (cytochrome c of diverse bird of prey species and vultures, graph modified after Seibold & Helbig 1995). The larger the genetic distance between homologous sequences of two species, the smaller is the portion of visible transition differences. Although statistically mutations causing transversions should occur more often than transitions (see Fig. 42), in closely related species the number of transitions is usually much higher. Obviously, mutations which conserve the basic chemical structure (purine or pyrimidine) occur more easily. When transition differences are replaced by transversion differences due to multiple substitutions, the proportion of transversions increases in time.

– Sequences of functional importance evolve more slowly and tend to conserve evolutionary novelties for longer intervals. These sequences have other disadvantages: the substitution rate varies with greater probability in an unpredictable way due to a correlation with changing environmental factors and population size, and substitution rates vary also within the gene depending on the functional importance of sequence regions.
– The visible differences between two sequences are called the uncorrected ***p***-distance, because the true number of substitutions may be higher than the visible one. This effect is the result of analogies and multiple substitutions (Fig. 165, more on this in ch. 8.2).

As systematists often analyse taxa separated by a long time, neutral sequences which get noisy quickly are not informative. Therefore, for studies on relationships between larger species groups usually coding sequences which are at least partly exposed to stabilizing selection are chosen.

When only some parts of a sequence evolve neutrally or are subject to weak selection while other regions are conserved, it is often assumed that the effect amounts to a slowing of the molecular clock when the average substitution rate of the whole sequence is considered. And it is often believed that over long periods of time sequence evolution is still a stochastic process. The consequences of this assumption for the practice of molecular systematics are not yet examined in detail. However, often sequence positions that evolve neutrally get rapidly noisy and therefore are not informative or even misleading for phylogenetic analyses, whereas in conserved positions none or only few substitutions occur so that the whole sequence is of little value. Spectral analysis is recommended to visualize the putative ***homology signal*** and contradicting noise contained in an alignment (ch. 6.5, Fig. 154).

Calibrating the molecular clock

To begin with, the evolutionary rate can only be determined as a relative value, namely the number of substitutions n per branch of a tree. This is obtained through comparison of the number of substitutions of terminal species of a given tree topology. The true number of substitutions n on a lineage (edge) of a dendrogram can be estimated with distance or maximum likelihood methods that correct for multiple (invisible) substitutions (see ch. 8.2. and 14.3). Assuming that the substitution rate λ is constant, the number of substitutions n on a branch is $\lambda \cdot t$, t being the unknown time between beginning and end of this branch. To determine the age of a monophylum using genetic distances, the time axis of a topology has to be "calibrated" and the relative branch lengths have to be transformed into absolute lengths with the help of an absolute substi-

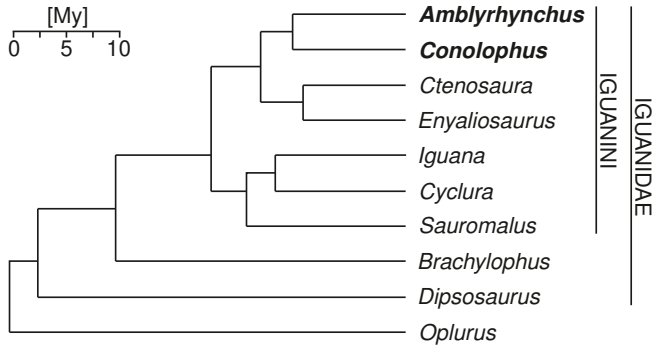


Fig. 40. An application of a molecular clock: analysis of the origin of the iguanas of the Galápagos *(Amblyrhynchus* and *Conolophus)*. Dendrogram reconstructed from ribosomal 12S and 16S genes of the mitochondrial DNA. Branch lengths represent divergence times. The relative branch lengths were obtained using a maximum likelihood analysis. To estimate the divergence time (see time scale; My = millions of years) the substitution rates known for ungulates were used (Rassmann 1997). The iguanas of the Galápagos are monophyletic. According to this estimate the divergence of the two genera (more than 10 million years) is older than the islands that are today above sea-level.

tution rate (substitutions per unit of time). To do this we need a dated point on the tree. This can be obtained from the fossil record or with dated biogeographic events. One can also use substitution rates known for the same gene but from other related species (Fig. 40).

Example: Nikoh et al. (1997) assume that the substitution rate of aldolase C is uniform in all eukaryotic organisms. The average genetic distance per position between recent amphibians and amniotes is about 0.17 substitutions in a dendrogram reconstructed with aldolase sequences. The fossil record gives a divergence time of about 350 million years separating recent amphibians and amniotes from their last common ancestor. The resulting substitution rate is $0.24 \cdot 10^{-9}$ substitutions per year and position (0.17/(2·350) my). Considering other pairs of eukaryote taxa, the average substitution rate is estimated to be $0.23 \cdot 10^{-9}$. – The distance between *Branchiostoma* and recent vertebrates is on average 0.36 substitutions per position. Using a rate of $0.23 \cdot 10^{-9}$ one gets a divergence time of 780 million years ($10^9 \cdot 0.36/(2 \cdot 0.23)$). (A more complex example with groups of different substitution rates can be read in Berbee et al. 1993). The quality of the estimation depends on the correct determination of the age and phylogenetic position of fossils and on a correct estimation of genetic distances (see ch. 8.2).

To count substitutions only terminal sequences can be compared, i.e. data obtainable for populations that exist today. Therefore we can only estimate average substitution values for the whole time separating two species, i.e. for the path which joins the sequences with the next basal node of the phylogenetic tree. An unnoticed irregularity of the substitution rate, i.e. an **irregular molecular clock** can lead to **wrong estimations** of divergence times, as illustrated schematically in Fig. 41. The reality is even more complex, because in phases of rapid evolution multiple substitutions (several substitutions at the same position) which are not noticed can accumulate, so that the genetic divergence is underestimated. Due to the errors that may occur estimating distances when variable positions are saturated by multiple substitutions it is important to work with sequences which (hopefully) are not saturated (saturation effects: see Figs. 39, 43).

In summary we note that neutrally evolving sequences indeed have the advantage that irregular variations of the speed of evolution should not be as frequent as in non-neutral sequences, but

– for the calculation of divergence times neutrality of substitutions is not as important as the constant ticking of the molecular clock or the correct estimation of the average substitution rate.

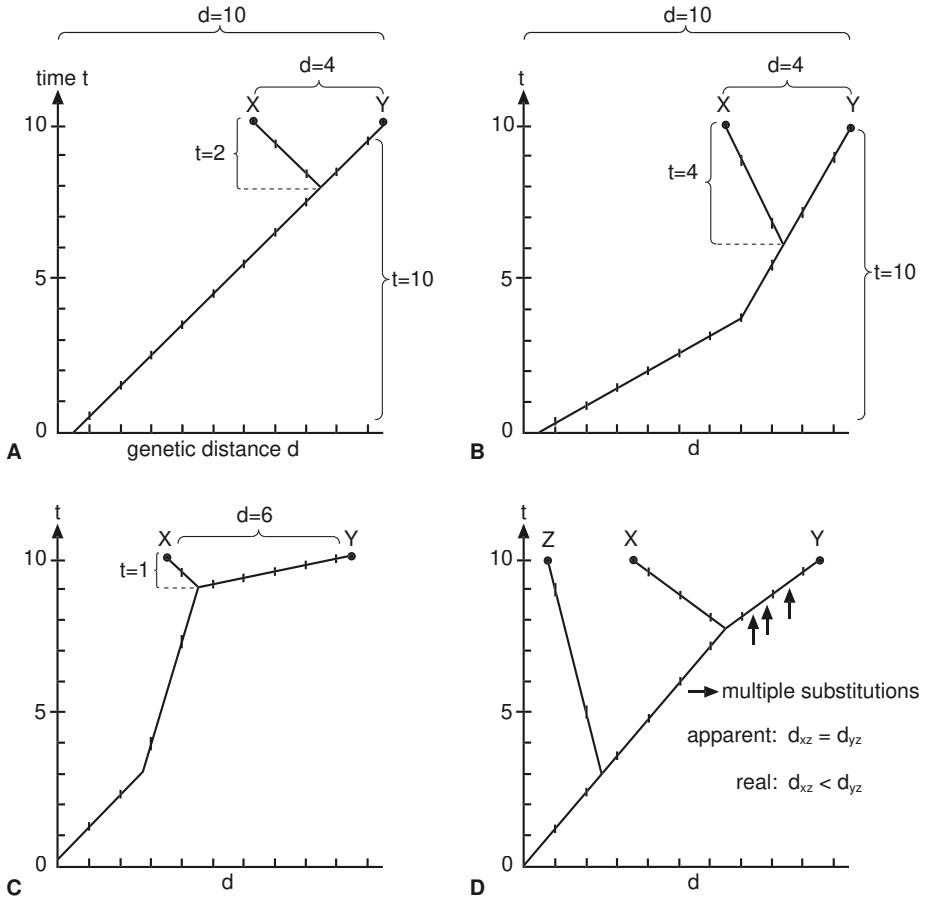


Fig. 41. Cases with an irregular molecular clock leading to a wrong estimation of divergence times. Each bar on the tree represents the same number of substitutions. **A:** With the existence of a steady clock (equivalent to a constant and universal substitution rate) the divergence time is $t = d/(2\lambda)$, whereby λ is the number of substitutions per unit of time (substitution rate). **B:** The substitution rate changed in a basal lineage. The sum of substitutions that occurred and thus the average of substitutions per unit of time is the same as in case A, but the age of the monophylum {X+Y} would be underestimated (t=2 instead of t=4) taking the rate from case A. **C:** The species Y shows many more substitutions than species X. This difference is detected in a comparison with a third species (see also "relative rates test", ch. 14.8). Without a correction for rate differences the age of monophylum {X+Y} would be estimated to be t=3. **D:** Multiple substitutions and irregular clocks cannot be detected in a comparison with a third species This situation can be expected when many multiple substitutions occur (more than one substitution at one sequence position) and sequences approach "saturation" (ch. 2.7.2.4).

Does the molecular clock tick irregularly?

The use of the "molecular clock" to date the age of monophyla (ch. 8) and to reconstruct phylogenetic relationships with simple distance methods requires the assumption that the clock ticks at a constant frequency, or that the average substitution rate of a sequence should always be the same along different lineages of a tree during thousands or millions of years. It has been explained in the preceding paragraph (Fig. 41) which mistakes crop up when the substitution rates fluctuate strongly. One has to assume

– that sequences of functional importance have varying evolutionary rates in a way comparable to phenotypic characters, with higher rates in phases of adaptive radiation (when selection pressure on variations is reduced), higher rates in periods of strong directional selec-

tion, lower ones during stabilizing selection (Fig. 41B).

Furthermore, there is the suspicion that the existence of a molecular clock is a hypothesis which often is not applicable even for neutrally evolving sequences, because factors independent of selection influence sequence evolution. Of particular importance is

- the dependence of genetic drift on population size (see Fig. 37).
- It is also conceivable that variations of generation times and of metabolic rates influence the mutation rate in the germ line.
- There may be differences in DNA repair efficiency,
- and variations in the exposure to mutagens.

Examples for variations of substitution rates: The substitution rate of rDNA-sequences probably was 20 times higher in the stem lineage of the Diptera than in other insects and later dropped by half (Friedrich & Tautz 1997). Lice of the genera *Geomydoecus* and *Thomomydoecus* are ectoparasites of pocket gophers (Geomyidae), with which they have coevolved. The COI-gene of the lice has substitution rates that are on average three times higher (ten times higher only for synonymous substitutions) than in the host species, an indication of the influence of the generation time on the substitution rate. – There are several examples for detectable episodic changes of substitution rates, while the causes are only the subject of speculations: the substitution rate in the mitochondrial genes of vertebrates is at least six times higher in mammals than in fish (influence of the metabolic rate), in 18S rDNA-sequences of parasitic angiosperms (Balanophoraceae, Hydnoraceae, Rafflesiaceae) it is 3.5 times higher than in autotrophic species (influence of population size); the same gene evolves 50 to 100 times faster in planktonic Foraminifera than in benthic species (influence of fluctuations of population size) (Adachi et al. 1993, Nickrent & Starr 1994, Pawlowski et al. 1997). – According to phylogenetic reconstructions, growth hormones of mammals evolved erratically, phases of rapid changes alternated with phases of few alterations (Wallis 1997). – There are also marked differences between closely related species and even among intraspecific groups. The substitution rate of house mice is markedly higher in populations of *Mus musculus domesticus* than in *Mus musculus musculus* (Boursot et al. 1996). In fruitflies (Drosophilidae) it was observed that the substitution rate of amino acids (enzyme GDPH) must have been 12 times higher in some stem lineages than in others (Kwiatowski et al. 1997). In hummingbirds (Trochilidae) a slowing down of the substitution rate in higher altitudes was detected (Bleiweiss 1997).

Several tests for the constancy of the molecular clock are available. These include the "relative rate test" (ch. 14.8), parametric bootstrapping (ch. 6.1.9.2), maximum likelihood methods (ch. 14.6.1) or edge-length tests (e.g., Takezaki et al. 1995, which has rarely been used). The relative rate test, which is based on distance data and model assumptions is useful to identify branches that evolve faster or slower than others, however, variations within a line that separates the last common ancestor of two taxa from a terminal taxon are not detected. Parametric bootstrapping relies on simulations of sequence evolution using model assumptions. In can be tested if in a simulation with a constant clock model a given topology is recovered. Within the framework of ML analyses entire phylogenies obtained with or without a constant clock model can be compared using a likelihood ratio test (explained in ch. 8.1).

In case rates are not constant and heterogeneous: new Bayesian approaches or ML local clock models allow the incorporation of rate heterogeneity into estimates of divergence time (Thorne et al. 1998, Yoder & Yang 2000, Aris-Brosou & Yang 2002, Thorne & Kishino 2002, Yang & Yoder 2003).

2.7.2.4 Evolutionary rates

Statements on evolutionary rates of molecules are mainly based on **pairwise comparisons of sequences** (ch. 8.2). It rarely is attempted to reconstruct a ground pattern of a gene for monophyla (e.g., Wheeler 2000) that allows estimate rates in stem lineages. In order to compare for example substitution rates of rodents with those of primates, pairwise comparisons of sequences with a not too distant outgroup (e.g., chicken) are necessary (mouse – chicken and chimpanzee – chicken). The difference in rates can then be attributed to the lines which lead from the last common ancestor of chimpanzee and mouse to the terminal taxa (see also "relative rate test", ch.

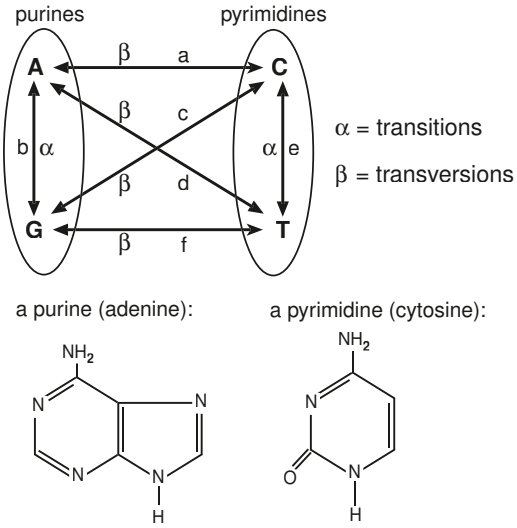


Fig. 42. Possible mutations of a sequence position; a-f are different types of substitutions discerned in models of sequence evolution.

14.8). In this way estimations for average substitution rates of long evolutionary lineages are obtained. The estimations are also often transformed with corrections which are based on further assumptions (see distance corrections, ch. 8 and ch. 14.3). Even though wrong or inexact assumptions may cause errors, clear tendencies arise from the wealth of known data which undoubtedly indicate the existence of lawful processes.

For good reasons the existence of a **universal "molecular clock"** is not seriously considered (see below). But all model-dependent methods of tree inference and molecular estimations of absolute ages of taxa (see ch. 2.7.2.3) rely on the existence of taxon-specific ephemeral or **"local molecular clocks"**.

If the assumption of the existence of a universal molecular clock were correct and therefore the evolutionary rate of single genes of different species would be the same, different species separated by the same divergence time from a common ancestor would have to show a similar number of substitutions for this period of time. An elegant proof for the **absence of a universal molecular clock** is the comparison of homologous genes of coevolving parasites or symbionts and their hosts. For example, in bacterial endosymbionts of the genus *Buchnera* living in aphids, substitution rates were observed that on average are 36 times higher than in their hosts (Moran et al. 1995), although the divergence time is the same. The molecular clock ticks differently in different organisms. Additionally to substitutions there are other mutations that occur at irregular intervals: if there were only substitutions, sequences would have to maintain the same length during evolution and spasmodic changes were only to be expected with irregular variations of the substitution rate. However, sequence lengths also vary substantially in nature. For example, the 18S rRNA gene is about 1800 nucleotides long in most eukaryotes, in several arthropods, however, it increased to more than 2400 nucleotides several times independently (i.e., in different lineages). Insertions or deletions can occur due to replication errors, leading to sudden, unpredictable changes in sequence length. Evolutionary rates for such events are not known, but obviously they are very irregular in the animal kingdom.

Evolutionary rates for nucleotide sequences vary, amongst others, depending on the selection pressure acting on certain functionally important sequence regions. Therefore genes evolve with different speeds. Also non-coding areas seem to be conserved to a varying degree when different organisms are compared. Furthermore, individual types of mutations do not occur with the same frequency. Well known is the analysis of the **transition/transversion rate (Ts/Tv)**, which specifies the relative frequency of these substitutions and is used in models of sequence evolution:

Transitions (point mutations: A⇔G or T⇔C), in which the chemical class of nucleotides is preserved, are less selected or originate more easily in the cell than **transversions** (point mutations: purine ⇔ pyrimidine) and therefore occur more frequently. As for each nucleotide, twice as many transversions are possible than transitions (see Fig. 42), the opposite should be expected (more frequent transversional substitutions), if all point mutations were equivalent. For example, in the mtDNA of hominids transition rates are about 17 times higher than transversion rates (Kondo et al. 1993). A consequence of the inequality of rates for transitions and transversions is that transitions are superimposed more rapidly by multiple substitutions than transversions (Fig. 43, 39). At first, transitions accumulate more rapidly, but then due to repeated mutations in the same positions

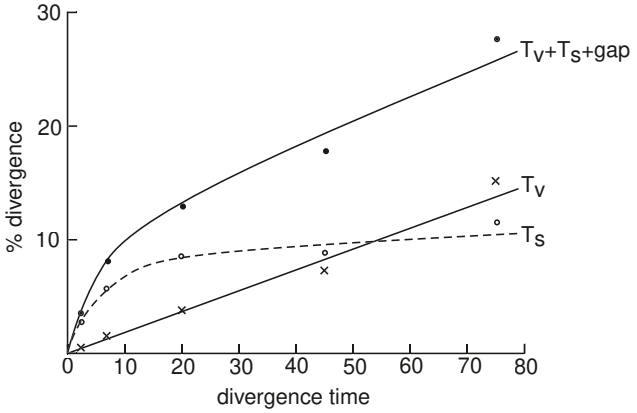


Fig. 43. Proportional divergence of mitochondrial sequences of a few mammals (ungulates, humans and mice) in pairwise sequence comparisons (after Miyamoto & Boyle 1989). Only the transversions (Tv) accumulate linearly with divergence time, the transitions (Ts) are saturated rapidly. This affects the whole divergence of the sequences (Tv + Ts + gaps).

the number of visible transitional substitutions decreases relative to the number of real substitution events.

The empirical observation of the Ts/Tv-rates is obstructed because in variable sequence positions transitions cause a more rapid accumulation of multiple substitutions than transversions. Thus two or more mutations can occur successively at one sequence position, so that in cases of longer divergence times the number of real substitutions is not directly visible. Estimating these invisible substitutions the **Ts/Tv-rate** has been calculated to be 5.5 on average for the cytochrome b gene of mammals, and 9.5 for the 12S rRNA gene, indicating that transitions are far more frequent than transversions. There are large **taxon-specific variations** (cytochrome b: about 1.0-18.6; 12S rDNA: 0.9-12.0; Purvis & Bromham 1997). A prerequisite for these determinations is the correct estimation of the divergence time. These data show that hypotheses based on the existence of uniform transversion rates for a comprehensive monophylum are probably wrong in most cases.

When comparing the same gene in different species that are separated by long divergence times the **visible genetic distance** of the sequences may not increase at the same rate as the **evolutionary distances** due to multiple substitutions. This phenomenon is called the "**saturation**" of a sequence: addition of substitutions to those positions that are free to vary will not increase the visible genetic distance. This term only refers to the information content of an alignment which is relevant from the point of view of systematists (Fig. 43); it is of methodological importance but does not imply biological effects.

Whether saturation occurs depends on the divergence time and on the speed of accumulation of substitutions. For this reason transitions are informative for shorter periods of time, while for longer periods transversions often contribute more information for a phylogenetic analysis (as long as they are also not saturated). When this phenomenon is not considered, estimated divergence times are too short. In distance analyses or in other modelling methods adequate corrections are incorporated to consider multiple substitutions (ch. 14.1.1). Furthermore one has to remember that in a sequence some of the positions can be highly conserved due to functional constraints, whereas neighbouring positions are variable. When numerous substitutions occur, they will mainly affect variable positions, where the events superimpose each other. To recollect: positions are variable, because they are functionally less important and therefore the selection pressure on mutated alleles is small.

Saturation can be detected with plots as in Fig. 43, but also indirectly with a phenomenological character analysis (see ch. 6.5), because multiple substitutions cause the signal-to-noise ratio to fall off. A problem are **hidden saturations**: if some

substitution	A → T	C → T *(Ts)*	T → C *(Ts)*	A → G *(Ts)*	G → A *(Ts)*	G → T	C → A
pseudogene	4.7 %	21.0 %	8.2 %	9.4 %	20.7 %	7.2 %	6.5 %
control region	0.4 %	25.8 %	33.8 %	14.1 %	20.0 %	1.1 %	1.1 %

Fig. 44. Example for proportions of specific substitutions in pseudogenes and for the control region of the mtDNA of mammals (after Gojobori et al. 1982, Li et al. 1984, Tamura & Nei 1993) *(Ts* = transitions).

taxon	Nematoda	Collembola	Hymenoptera	Echinoidea	Mammalia
AT-content (%)	69-74	60-71	76-80	57-59	58-61.5

Fig. 45. AT-content of the COII-gene of some animals (from Simon et al. 1994). (The AT-content is the proportion of the bases adenine plus thymine of all nucleotides of a sequence).

gene regions are highly conserved and a few positions are variable, one gets low total distance values comparing two sequences. This may be misleading, because the few variable positions may very well be saturated.

Also other types of substitutions do not occur homogeneously. For example, in pseudogenes of mammals different proportions (expressed as percent of the total number of substitutions) were observed (Fig. 44).

The values in Fig. 44 for the pseudogenes and those for the control region are not directly comparable, because the values of the lower line come from the same species *(Homo sapiens)*, whereas the values for the pseudogenes represent the average for 13 different mammal species. Therefore, the divergence of the sequences examined is greater for the pseudogenes. It is conspicuous that the direction of a substitution influences the rate (compare C → T and T → C, A → G and G → A). Which mechanisms lead to these variations will not be examined here (for details see e.g., Li 1997), however systematists should know these irregularities exist. The assumption that substitution rates are uniform for all nucleotides probably is rarely correct. When it is used for modelling of sequence evolution, it is a source of error.

Further irregularities exist in the **shift of nucleotide frequencies** from equal distribution (1:1:1:1) to an accumulation of A-T or G-C base-pairings. Nucleotide frequencies vary from taxon to taxon and therefore also the probability for the occurrence of specific substitution types is not always the same (Fig. 45).

Variations of substitution rates also exist between larger gene regions. Comparison of the secondary structure of **rRNA**-molecules has shown that some (not all) areas in loops have a higher variability than double-stranded areas. This is not surprising because a mutation in a helical area can interrupt the hydrogen bonds between base pairs. However, a universal rule cannot be derived from this observation: there are stem regions (with paired nucleotides) which are very variable and single stranded loops in which only rarely substitutions are seen (Fig. 46). In **t-RNA**-sequences just the anticodon-loops are conserved. The explanation for these variations is the different selection pressure, which depends only on the functional significance of molecular regions. Helical regions determine to a large degree the molecule's tertiary structure; several sequence regions are important for the binding of ribosomal proteins, of mRNA or tRNA, and therefore are highly conserved. Wherever the three-dimensional form of the molecule is important, mutations are probably selected or compensated by matching mutations in the complementary strand. (Remember: as a rule alignments do not justify conclusions on the frequency of individual mutations; only those mutations in an alignment which spread in the population and which were conserved, i.e. mutations that became substitutions and that were not masked by subsequent substitutions are visible.)

Introns of protein coding sequences often show the same substitution rate as synonymous substitutions of exons, which is in accordance with the assumption of neutral evolution. (Remember: introns are sequence areas that do not code for RNA or proteins. They are inserted in gene sequences. The coding sections are called exons). For example, in rodents synonymous substitutions in exons and substitutions in introns are three times more frequent than non-synonymous

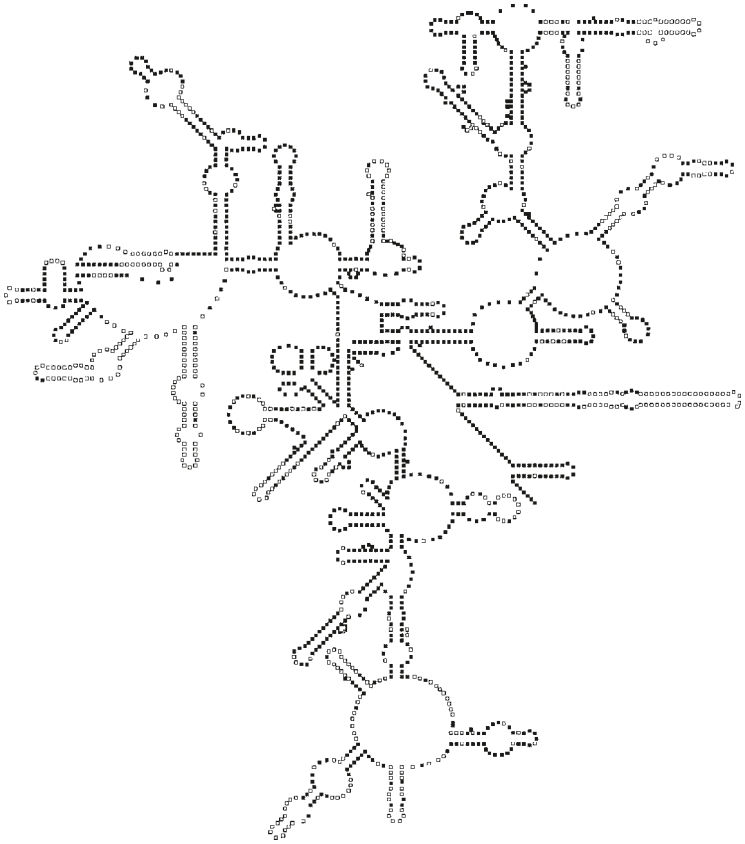


Fig. 46. Variability of the substitution rate for individual positions of the 18S rRNA gene of eukaryotes, plotted on the reconstructed secondary structure of the molecule of *Saccharomyces cerevisiae*. Conserved positions are marked in black, variable ones in white (modified after van de Peer 1997, see also van de Peer et al. 1996).

substitutions (Hughes & Yeager 1997). **Non-coding sequences** are apparently without function, but one cannot assume that they always evolve neutrally. Non-neutral evolution occurs, for example, in ITS-sequences (see below). It is recommendable not to presuppose a lack of selection pressure without testing.

Sequences evolving rapidly or "neutrally" are particularly interesting for population studies. The control region of the mtDNA of vertebrates is often sequenced for this purpose. As the control regions of invertebrates are often extremely rich in A and T as well as very variable in length (Crozier & Crozier 1993), they are less suited for phylogenetic studies in this case, because it is difficult or impossible to find the optimal alignment with homologous nucleotides in columns (alignment procedures: see ch. 5.2.2.1).

ITS-sequences (internal transcribed spacers) that separate ribosomal genes belong to the non-coding sequences, which are assumed to evolve neutrally. Therefore they have been sequenced for studies of population genetics. There are however observations suggesting that at least the secondary structure and partially also the sequence is highly conserved in some taxa. This indicates some function of the secondary structure in the processing of the primary RNA-transcript (e.g., Mai & Coleman 1997).

Amino acid sequences also evolve at different rates. Though mutations occur at the level of DNA molecules, selection operates at the level of functional units that influence the organism's fitness. Therefore rates of protein evolution depend on the functional importance of protein structure. Histones nearly do not change, hemo-

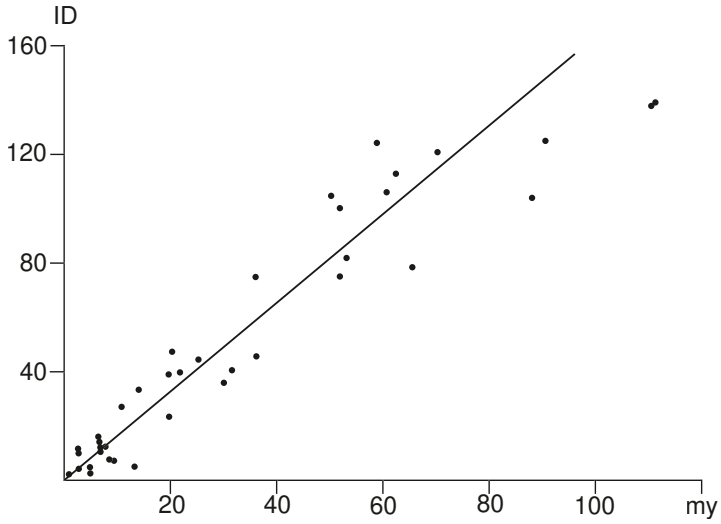


Fig. 47. Immunological distances (**ID**) of albumins and divergence time of diverse vertebrates (Anura, crocodiles, ungulates, carnivores, after Maxon 1992). The divergence time has been verified by comparison with the fossil record. It is obvious that the immunological distance is a good measure for the estimation of the divergence time, an observation that can be attributed to the neutral evolution of these proteins.

globins change slowly, immunoglobulins more rapidly. For proteins of mammals mean substitution rates of about 0.7 substitutions per sequence position in 10^9 years were estimated (Li 1997). In case the proteins have not been sequenced, the level of conservation can be examined studying the variability of the corresponding genes. For the mitochondrial genome it is known that NADH-genes vary markedly stronger than cytochrome genes. This has to be explained with the higher functional significance of the tertiary structure of cytochromes. Similarly, individual regions of molecules evolve at different rates: for cytochrome b proteins and cytochrome oxidase it has been shown that external areas of the folded molecule, which are important for the enzymatic reaction, are less variable than regions which are positioned in the membrane of the organelle or in the mitochondrial matrix (Irwin et. al 1991, Disotell et al. 1992; further examples in Kimura & Ohta 1973, Kimura 1983, Green & Chambon 1986).

The similarity of **serum albumins** of vertebrates has often been examined with immunological methods. As albumins have a relatively unspecific function (osmoregulation, protein reserve, transport functions), the assumptions seem to be justified that they evolve "neutrally" (independent of natural selection) and that strong variations of evolutionary rates caused by the influence of environmental factors are not to be expected (which has often but not always been verified with analyses of different taxa of vertebrates: Cadle 1988, Maxon 1992). Based on these prerequisites, immunological distances between species can be interpreted as a measure for the divergence time. A reference to the number of nucleotide substitutions in coding DNA cannot be unequivocally established, but the plausibility of the immunologically inferred phylogenetic relationships can be confirmed by comparison with the fossil record and with geological events (e.g., Joger 1996).

The comparison of **mitochondrial** and nuclear genes (e.g., cytochrome oxidase, rDNA) of the same species shows that mitochondrial genes are more variable. This is ascribed among others to the higher metabolic rate of mitochondria, causing a higher mutation rate. The rate for synonymous positions of homologous genes of mammals is about ten times higher in the mitochondria than in the nucleus, whereas the genome size and gene order change little. Additionally, the rates vary from taxon to taxon: the evolution rate of the third codon position of mitochondrial genes is said to be three times higher in species of *Drosophila* than in mammals (Sharp & Li 1989).

aminoacid	Ala	Ala	Ala	Ala	As	As	Gly	Gly	Gly	Gly	Trp
1. codon position	G	G	G	G	G	G	G	G	G	G	U
2. codon position	C	C	C	C	A	A	G	G	G	G	G
3. codon position	***A***	***C***	***G***	***U***	***C***	***U***	***A***	***C***	***G***	***U***	**G**

Fig. 48. Example of the degeneration of the universal genetic code. Italics: variable positions. Codons coding for the same amino acid are called "synonymous". Synonymous substitutions mostly occur in the third codon position.

Since mitochondria are not directly exposed to environmental conditions as are phenotypic characters, it could be assumed that mitochondrial genes evolve stochastically and with comparable rates in different species. Empirical data, however, prove that this assumption often is not true. The accelerated evolution of COII in primates and the irregular distribution of the frequency of synonymous to non-synonymous substitutions in populations of closely related species indicate that an influence of selection or consequences of genetic drift exist (Ballard & Kreitman 1995). Furthermore, it could be expected that the evolutionary rate of mitochondrial genes is independent of the generation time of species, because the multiplication of mitochondria is not synchronized with the production of germ cells. Nevertheless, a correlation with generation time has been found in comparing different species (Bromham et al. 1996).

In alignments of **protein-coding DNA sequences** it can be seen that the **third codon position** varies more strongly than the first (e.g., Sharp & Li 1989). The cause is again the selection pressure: many mutations of the third codon position (ca. 70 %) have no effect on the coded amino acid (synonymous substitution), because of the "degeneration" of the genetic code (Kimura 1983; see Fig. 48). For 20 amino acids, 64 (4^3) possible codons are available (which are not always used in the same way in all organisms). **Synonymous substitutions**, meaning substitutions which do not change the coded amino acid, are more frequent: the substitution rates can reach tenfold in comparison with non-synonymous substitutions, although statistically (without selection effects) more non-synonymous mutations should occur. As the third position gets noisy rapidly due to multiple substitutions, it is often excluded from phylogenetic analyses. Instead of using the translated amino acid sequence, exclusion of the third codon position decreases the variability in an alignment in a similar way, but more characters are retained. It is also recommendable to encode aligned codon positions with many synonymous substitutions in the R-Y-alphabet, in order to take into account only transversions. This may help to reduce the noise in the data.

It does not always seem to be trivial which nucleotide is at the third position, because nucleotides are not distributed as evenly as they should: all codons which code for the same amino acid should occur with the same frequency in the genome. However, this is not the case. Depending on the organism, obviously other codons are favoured for the same amino acid (Grantham et al. 1990). There exists a certain selection pressure even for synonymous mutations (Bulmer 1988). An explanation for this phenomenon is the supposed preference of those codons for which more tRNA-molecules are available in the cell or which are generally transcribed more efficiently.

Transitions and transversions have different effects on proteins. On average transversions seem to induce more drastic alterations and cause more often nonsynonymous changes (analyses of mammalian genes: Zhang 2000). The resulting amino acid substitutions can be classified as either conservative (no dramatic change of physicochemical properties of the amino acid) or radical. Radical substitutions occur with a slightly lower rate than conservative ones, they seem to evolve under a somewhat stronger purifying selection (Zhang 2000), and they are caused more frequently by transversions than by transitions.

A suggestion for differential weighting of the probability of homology of substitutions (ch. 6.3.1) is the classification of coding positions according to whether 4, 2 or 0 nucleotides can be substituted synonymously. This, however, requires that all synonymous codons are equally frequent in an organism, which in nature often is not the case.

The ratio of the usage of individual codons is estimated with the RSCU-value (RSCU = relative synonymous codon usage; Sharp et al. 1986):

X_{ij} is the frequency of the codon **j** for the amino acid **i**, and **n** is the number of alternative codons for the amino acid **i**. The ratio is given by the formula

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

A further, analogous possibility to evaluate substitutions is the study of the frequency with which amino acids are exchanged. Tables containing specifications on the frequency of specific substitutions (see Dayhoff et al. 1978) can be used for the weighting of putative homologies, when one chooses to evaluate the probability of events (ch. 6.3.1, 8.2.3).

The previously mentioned cases should make clear that there are all kinds of transitions between the extremes of neutral and of completely conserved DNA-regions. The theory of neutral evolution offers an explanation for these differences: some mutations are selected more strongly than others.

As already mentioned above, additionally to the variations in the evolutionary rates detected when different molecules are compared, differences can also be found when **comparing species**. For example, mitochondrial genes of sharks evolve slower than those of mammals; comparing rodents with other mammals it was noted that some proteins in rodents and other ones in mammals show large variations in substitution rates (Gu & Li 1992). The accelerated evolutionary rate of the mitochondrial protein COII in primates indicates that a selection process has been at work. A comparison of several genes showed that non-synonymous substitutions per position and unit of time occurred about twice as often in rodents than in other mammals (Ohta 1995): the higher rate in Rodentia can be explained by the shorter generation times (Wu & Li 1985). In insects again higher rates were discovered than in mammals (Sharp & Li 1989).

Several factors have been considered to be the causes for **taxon-specific** or **life-form specific evolutionary rates** (compare examples in ch. 2.7.2.3), for which, however, only a few correlation analyses have been done so far to prove their effect (Bromham et al. 1996):

- higher physiological rates could cause more damage to the DNA through radicals. This could explain the higher substitution rates in mitochondria in comparison to homologous nuclear genes, as well as faster evolution in endothermic species,
- polymerases could work with varying precision,
- the efficiency of repair mechanisms and the precision of DNA-replication could vary,
- the generation time or the number of cell divisions in the germ line could influence the mutation rate; this could explain, for example, why rodents evolve faster than hominids and lice of animals faster than their hosts. In the same way body size could correlate negatively with the substitution rate,
- the selection pressure on an allele can vary from population to population depending on the environmental conditions,
- the population size has an influence on the genetic drift; therefore species with small populations (e.g., parasitic angiosperms) change faster.

Variations in the evolutionary rates exist when comparing

- coding and non-coding sequences,
- codon positions 1, 2 and 3,
- synonymous and non-synonymous codons,
- synonymous codons in mitochondria and nuclei,
- different regions of a molecule,
- homologous regions of a molecule in different taxa,
- mitochondrial and nuclear genes.

In general, rates of molecular evolution can vary within and between lineages

- due to changes in mutation rates,
- changes of population size,
- and changes of selection effects (Bromham & Penny 2003).

Molecular systematists often axiomatically assume that the course of sequence evolution is a stochastic process. However, to rely on such an axiom is a risky enterprise since due to the influence of population fluctuations and unknown selection effects unforeseeable and episodic changes in substitution rates are to be expected (see ch. 2.7.2.1).

2.8 Summary: Constructs, processes and systems

The following table is not complete and is meant to stimulate reflection. "Things" and "material systems" are individual objects that really exist outside our minds. In the same way, the terms for processes refer to individual processes that occur in nature. Each term itself is a construct already. Each abstraction is a hypothesis and requires a thorough scrutiny of its quality.

things / material systems / state of material things	**processes / events**	**constructs**
individual organism	reproduction	species
individual organs and other material structures	inheritance	monophyla, stem lineages
	horizontal gene transfer	taxa
individual molecules	mutation and substitution	characters
functional reproductive communities	development of a reproductive barrier	homologies apomorphies
properties of organisms or of parts of organisms	genetic drift	genetic information
	selection, evolutionary adaptation	concepts for the term 'population'
	increase of genetic differences between groups of organisms (genetic divergence)	the system of organisms dendrograms

Some terms can either refer to constructs or to real existing material objects. Whether a "population" is a real system or the mental grouping of objects has to be decided for each individual case (see ch. 2.2). For all constructs (see ch. 1.2) scientists have to ask which relation they have to the material extrasubjective nature and whether the corresponding terms are useful tools for science.

The subjects "selection", "adaptation", "origin of reproductive isolation" and other main areas of evolutionary research will not be elucidated further in this book, they are treated in other textbooks. Despite their significance for the detection of mechanisms that cause speciation and phylogeny, they supply only a few arguments for the practice of systematization.

3. Phylogenetic graphs

3.1 Ontology and terms

Ontology asks what type of reality (the "being") is represented with the words of our language or with our concepts. We have to be aware of whether the technical terms and names we use in systematics refer to material things or to (mental) constructs, because the latter are subject to discussion and could be conceived differently, whereas material things are as they are, independent of any subject. In the following chapters this question has to be asked repeatedly.

Always remember that phylogenetic trees (= **dendrogams**, phylograms, cladograms, tree graphs) only visualize hypotheses and that the species or groups of species represented therein with symbols are only concepts of methodological importance (see ch. 2.6. and 2.3).

We can differentiate the following parts in a phylogenetic tree (Fig. 49):

- each line (= edge) represents a continuum of ancestors and their descendants, following consecutively each other along the time axis, and being parts of a clonal population or of a reproductive community at each time horizon, independently of whether along a 'lineage' several chrono- or other species shall be distinguished or not.
- Internal branches (= inner edges, internodes) always represent stem lineages; these are groups of organisms belonging to one or several consecutive species.
- Terminal branches represent single species, which are extinct or with still existing recent populations, or they represent stem lineages of terminal monophyla.
- Nodes (vertices) symbolize speciation events; these are processes occurring at the level of populations. The points can also be considered as symbols for stem species, stem populations or ground patterns. The interpretation often results from the context of the corresponding text.

Each internal edge of a diagram with dicho- and/or polytomies (a diagram without networks) separates two groups of taxa. Such a bipartition is also called a "**split**". Depending on the mode of illustration, the **edge length** represents the genetic distance, the number of character changes, a measure for the probability that the data support this edge, or only the separation of the two groups. Characters which substantiate a split or that supported it in some calculation are **split-supporting characters**.

Attention: in most graphs the edge length does not correspond to the divergence time. In many diagrams which are to represent a real character evolution, the edge length is the product λt of the rate of change λ of the character (substitution rate) and the time t. Small rates and long time periods as well as high rates and short time periods can yield similar long edges.

Dendrograms are usually depicted as dichotomous diagrams, implying that after a speciation exactly two reproductively isolated or irreversibly diverging populations are present. This seems to be the usual case in nature. The occurrence of multiple speciations, which can be illustrated as polytomy, can, however, not be ruled out (see ch. 2.5.2). In each case the dendrogram represents the state of knowledge of an author, or the result of a calculation, not necessarily the real natural history.

In graph theory dichotomous dendrograms are "binary trees", in which each **node** (= **vertex**) is linked to not more than three other nodes. Internal nodes or branching points of a phylogenetic tree represent ancestors or ground patterns, terminal nodes or "leaves" are terminal taxa which often represent recent species.

Each assemblage of organisms named with proper names by systematists is a **taxon** (see ch. 3.5). If one wants a taxon to be accepted by other systematists, it should represent a monophylum ("be monophyletic"). Each branch cut off the dendrogram represents together with all its attached

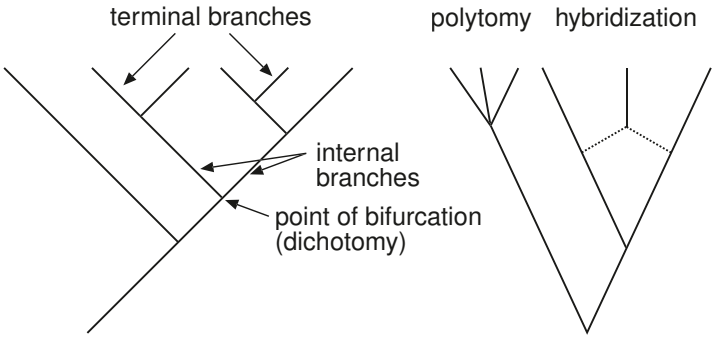


Fig. 49. Terms used to name sections of topologies.

smaller branches a putative monophylum whenever the dendrogram visualizes a hypothesis of phylogeny. We differentiate between monophylum and taxon, because it is impossible and undesirable to name each monophylum: there are many more monophyla than taxa (see ch. 2.6 and 3.4). Also, a monophyletic taxon is always a logical class and not a real material individual, because it is not an object or material system existing in a moment in time. The distinction and naming of taxa is based on conventions (see Fig. 24). Therefore the group which is to be referred to with a specific **taxon name** has to be defined (established) (Mahner & Bunge 1997). The definition refers to a point of divergence in time (= **systematizing definition**, see ch. 2.6, ch. 4.4) or to characters which can be derived from real properties of organisms (= **classifying definition**). With both methods boundaries of monophyletic taxa can be determined. When a monophylum has been recognized, the systematist has to estimate whether this monophylum will be frequently

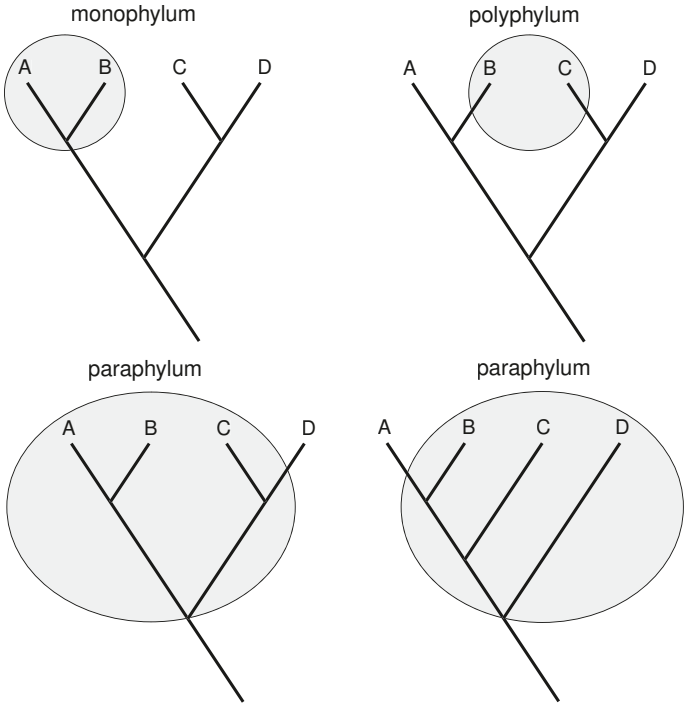


Fig. 50. Possible groupings of terminal taxa.

mentioned in practice and if it will also be identified by other scientists. In this case a **proper name** could be useful for communication between scientists. The fact that a monophylum is distinguished in such a way is a convention. Attention: a definition refers to the equation of a taxon name with a monophylum. Monophyla on the other hand are not defined, they are identified or discovered.

A **terminal taxon** (**OTU** "operational taxonomic unit") is represented in a dendrogram by a point at the end of a terminal edge. It symbolizes a group of organisms which are assumed to be monophyletic. Shall the graphic visualize a reconstruction of phylogeny, the corresponding characters of a terminal taxon used in the data matrix have to be **ground pattern characters** of the taxon. These are characters which are assumed to have been present in the last common ancestor of the members of the taxon (ch. 5.3.2).

In a dendrogram groups are differentiated according to their descent and their composition (Fig. 50).

Distinguishing features of these groups are:

- **Monophyletic groups** contain all descendants of a single stem species (or of a stem population or of a single ancestor). The stem species is included in the monophylum (ch. 2.6).
- **Polyphyletic groups** contain species derived from different stem lineages, and they do not include all descendants of the last common ancestor. According to Hennig (1966, 1982), polyphyly is supported by analogies and parallelisms (see ch. 4.2.3). In trees derived from sequence data polyphyly may have different causes.
- **Paraphyletic groups** contain only some and not all of the descendants of a single stem lineage and thus never form a terminal taxon. According to Hennig (1966, 1982), paraphyly is supported by plesiomorphies (see ch. 4.2.3). In trees derived from sequence data paraphyly may have different causes.

Topology: the relative position of taxa to each other in a rooted or unrooted dendrogram.

Edge: line separating taxa or groups of taxa in a dendrogram, also called a "branch".

Inner edge: edge separating two groups, each composed of two or more taxa .

Inner node: junction of 3 or more edges (= branching point).

Terminal taxon: taxon connected only to one edge; it can be a species without daughter species or a monophylum represented by its ground pattern or by its stem species.

Split: bipartition of a set of species.

3.2 Topology

The term **topology** refers to the relative position or spatial relation of taxa to each other in a tree graph, which is independent of the type of geometric illustration, of the spatial perspective, and of the position of the "root". An unrooted dendrogram can show the same relative positions of taxa as a phylogenetic tree (see Fig. 51, 52).

3.2.1 Visualization of compatible hypotheses of monophyly

Each phylogenetic tree is a summary of several hypotheses of monophyly visualized in a clear diagram. The hypotheses are compatible or congruent whenever they can be combined into a single Venn diagram without any intersections (Fig. 51).

There exist no conventions (and even no arguments in favour of their introduction) on how to depict phylogenetic trees. They can be rectangular, slanted or radial, drawn as a Venn diagram or written as a formula with parentheses. A hypothesis is described with the topology and the position of the root. A specific **topology** contains a specific relative position of the nodes and of the terminal taxa to each other, independent of the position of the root (the origin) and the form of the graphic illustration. The lines between the nodes are the edges. Edges can represent stem lineages or only the presence of characters in the

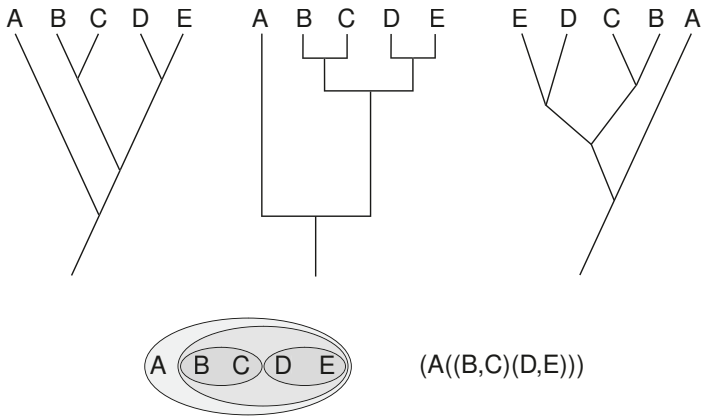


Fig. 51. Equivalent illustrations of a topology. Top: tree graphs; lower left: Venn diagram; lower right: bracket diagram.

analysed dataset when the topology does not represent a hypothesis of phylogeny.

– The topology of a dichotomous tree is unambiguously defined when for each inner point (node) of the diagram three neighbouring points are determined.

The orientation in space and the angles between edges are irrelevant here; branches can be rotated in any direction at each point without modifying the topology. Whenever the length of the edges is not thought to represent distances or the number of supporting characters, their length ratio to each other is irrelevant. There is a tradition to name a dendrogram with branches drawn to scale a **phylogram**.

Several programs safe trees in form of tree files. These show the encaptic order* of taxa with parentheses and they may also include information on branch lengths, as for example:

"(((species1:0.324, species2:0.300),species3:0.432), species4:0.511);".

If opened with tree drawing software the correct branch lengths will be shown in the phylogram or in a radial tree (note: the final semicolon may be required by some programs).

An "unrooted" or unpolarized diagram contains less information than a dendrogram. In Fig. 51 it can be seen that taxon A is located at the base and that it is the adelphotaxon to (B,C,D,E). The unrooted diagram (Fig. 52) shows the same topology. However, the point of origin is lacking and therefore also one node is missing, with the result that the relative age of taxon A is not seen.

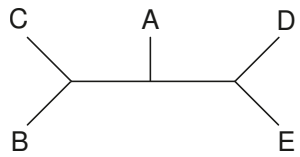


Fig. 52. Unrooted (unpolarized) or radial topology.

Unrooted diagrams are not "worthless", because they contain all groupings which are also present in the corresponding phylogenetic tree, but they are poorer in information. The polarity of characters is not visible, the graph does not allow the reconstruction of evolution.

If no phylogenetic information is present in a dataset, this fact can be visualized with a "**bush diagram**" (Fig. 53). The best illustrations of contradicting information, showing incompatible groupings of taxa, are Venn diagrams and network diagrams (see Figs. 54, 55).

* The term "encaptic" is often used in German phylogenetic literature. It implies that one group is encapsulated in another one and emphasizes the unique position within a hierarchy excluding other memberships.

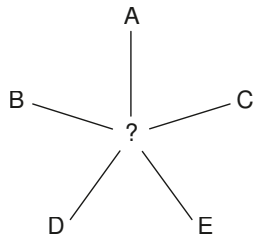


Fig. 53. Bush- or star-like topology. The question mark indicates that phylogenetic relationships are not known.

Furthermore, topologies have to be distinguished which do not represent genetic divergences from those in which distances are drawn to scale. In scaled graphics, for example, the relative length of lines may correspond to the number of estimated substitutions which occurred within a stem lineage. This illustration is especially chosen for distance trees (ch. 8.2).

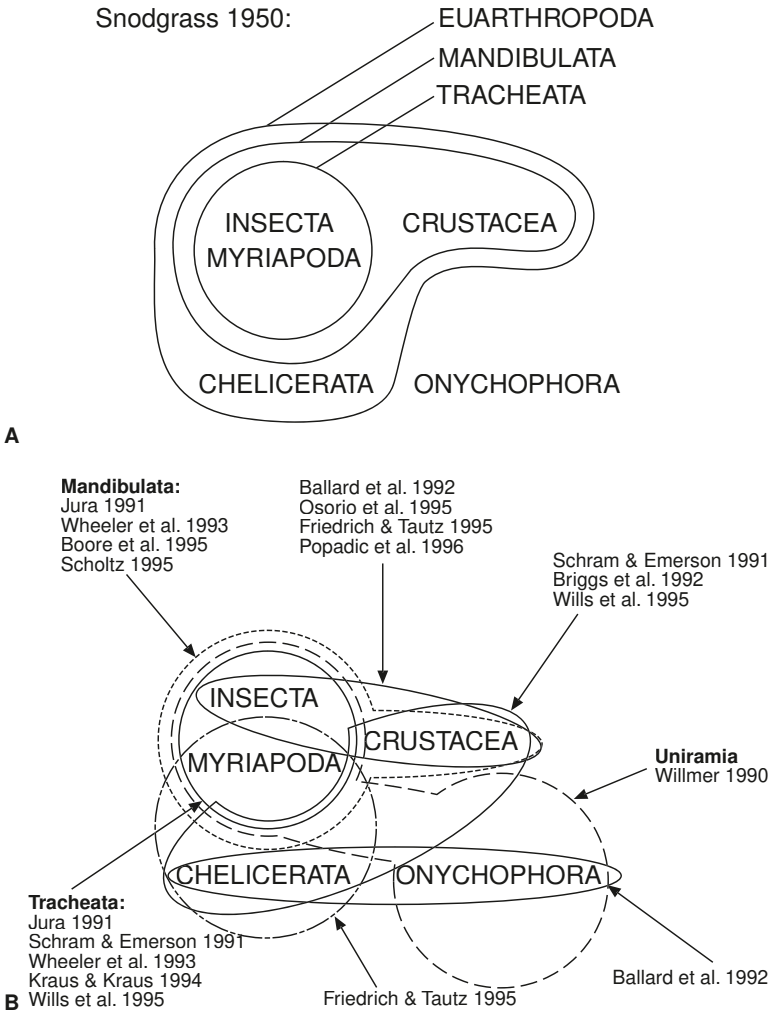


Fig. 54. Incompatible hypotheses on the phylogeny of arthropods (Venn diagram). **A.** According to Snodgrass (1950) the Euarthropoda, Mandibulata and Tracheata form an encaptic (hierarchical) order. The taxa are compatible with each other, the monophyly of Chelicerata, Crustacea, etc. is presupposed. **B.** Some results of new analyses are neither compatible with the traditional phylogeny nor with each other. Therefore several or in the worst case all of them have to be erroneous. The fact that in A no incompatible groups are visible does not mean that the hypotheses are correct, but only that no contradictions are indicated in the graph.

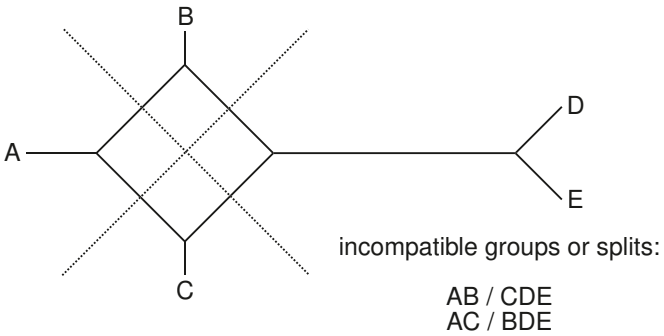


Fig. 55. Network diagram with incompatible splits. Some characters may support the split {A, B / C, D, E}, other ones the split {A, C / B, D, E}. This indicates that hypotheses (homology of characters) are inconsistent.

3.2.2 Visualization of incompatible hypotheses of monophyly

Whenever hypotheses on the monophyly of groups of species cannot be depicted without intersections in a Venn diagram (Fig. 54), the overlapping groups represent incompatible hypotheses of monophyly. In these cases it has to be considered that each hypothesis of monophyly always consists of two groups (in-group/outgroup), and therefore a statement of monophyly also implies a statement concerning the composition of other groups.

Incompatible hypotheses of monophyly can also be depicted as network diagrams (Fig. 55). If the supporting homology hypotheses are plotted on the corresponding edges, information on the incompatibility of putative homologies is visualized.

Network diagrams have the advantage that several alternatives can be visualized in one graph. When the length of the edges is representing a measure for the support of a split or e.g., a measure for the genetic distance, the proportion of the support for alternative hypothesis becomes visible quantitatively (e.g., split decomposition, ch. 6.4 and 14.4).

Another possibility is the illustration of one dendrogram for each alternative hypothesis. Such alternatives can be summarized with consensus diagrams (see ch. 3.3). These, however, contain less information than network diagrams.

3.2.3 Visualization of hypotheses of character polarity and of apomorphy

It can be shown in a graph which characters or character states have been evaluated to be apomorphies for a group of species, by writing a symbol corresponding to a novelty at the edge that represents the group's stem line (see e.g., Fig. 57, 76, 104, 152). In unrooted diagrams this only depicts character state changes or the occurrence of novelties, however without determining a polarity. One could not see, for example, if the novelty is the reduction or, alternatively, the

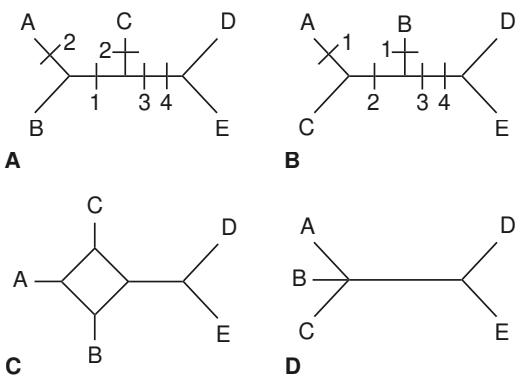


Fig. 56. Scheme to illustrate differences between network diagrams and consensus diagrams. The network diagram contains more information. **A** and **B** are topologies supported by a given dataset, **C** is the corresponding network graph, **D** a consensus topology. The latter does not indicate that there exists information in favour for the group (A+C) and also in favour of (A+B). The same consensus would also result in case there is no information for the grouping of the taxa A, B, and C. The numbers specify the characters of the given dataset that change their state along an edge.

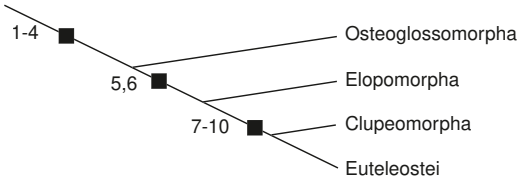


Fig. 57. Example for an argumentation scheme: phylogeny of the modern bony fishes (Teleostei) according to Lauder & Liem (1983). The numbers written at the edges of the dendrogram represent character state changes or symbolize the origin of evolutionary novelties (= apomorphies) and also imply an interpretation of character polarity. The apomorphies for this case are: 1: Presence of an endoskeletal basihyale. 2: Four pairs of pharyngobranchials present. 3: Three hypobranchials present. 4: Basibranchial and basihyal cartilages overlain by median tooth plates. 5: Two uroneurals extend anteriorly over the second ural centre. 6: Epipleural intermuscular bones developed throughout the abdominal and anterior caudal regions. 7: Retroarticular bone excluded from the quadromandibular joint surface. 8: Tooth plates fused with endoskeletal gill arch elements. 9: Neural arch on ural centre one reduced or absent. 10: Particular bone co-ossified with the angular. (Attention: this argumentation is incomplete as long as it is not discussed why these characters are homologous, and which are the plesiomorphic character states, and which are the autapomorphies of the terminal taxa, and why the latter are monophyletic. These arguments must appear in the text that accompanies the argumentation scheme.)

emergence of an organ. In rooted diagrams the novelty is an apomorphy of the following younger node. Accordingly, the characters written at an edge are characters of the ground pattern of the following monophylum. The ground pattern is symbolized with the younger node of the edge, and the node can be understood to represent the stem population (or the stem individual in other cases) of a monophylum. The order in which characters are listed at an edge is arbitrarily chosen and does not correspond to the sequence of the evolution of characters, unless this is explicitly stated.

The arguments in favour of a hypothesized phylogeny summarized in Fig. 57 form together with the corresponding list and discussion of characters an **argumentation scheme**. This may also contain contradicting characters (e.g., convergences, see ch. 4.2). In chapters 5 and 7 it will be discussed how characters can be evaluated phenomenologically or with modelling methods, and in chapters 6 and 8 the corresponding methods of phylogeny inference will be presented.

3.3 Consensus dendrograms

When several equally well founded but partly incompatible dendrograms are reconstructed with the same method and using the same set of data and when no additional information is available that justifies the selection of one of these dendrograms as a basis for a hypothesis of relationships, a consensus dendrogram can be used to visualize which relationships remain undisputed and which nodes are unresolved. It is also recommended to illustrate the result of bootstrap and jackknifing tests (ch. 6.1.9.2) with consensus trees. (An alternative to visualize conflicts are networks estimated from the data (ch. 6.4), which unfortunately are not frequently used).

A consensus dendrogram is not calculated from original data, but from already existing dendrograms. The topologies used for the consensus can be obtained from a single dataset, for example, when it is intended to eliminate the contradictions of alternative optimal topologies and to keep congruent parts, or from different data and different analyses (e.g., of morphological and molecular data) of the same set of species to summarize and visualize the congruent parts. A disadvantage of consensus trees is that relative branch lengths cannot be calculated using the currently popular maximum parsimony methods (but see also Bayesian analyses (ch. 8.4), where branch lengths are available).

Different methods to construct consensus dendrograms have been proposed:

Strict consensus method: of the compared dendrograms only those groupings are taken that occur in all topologies (Fig. 58).

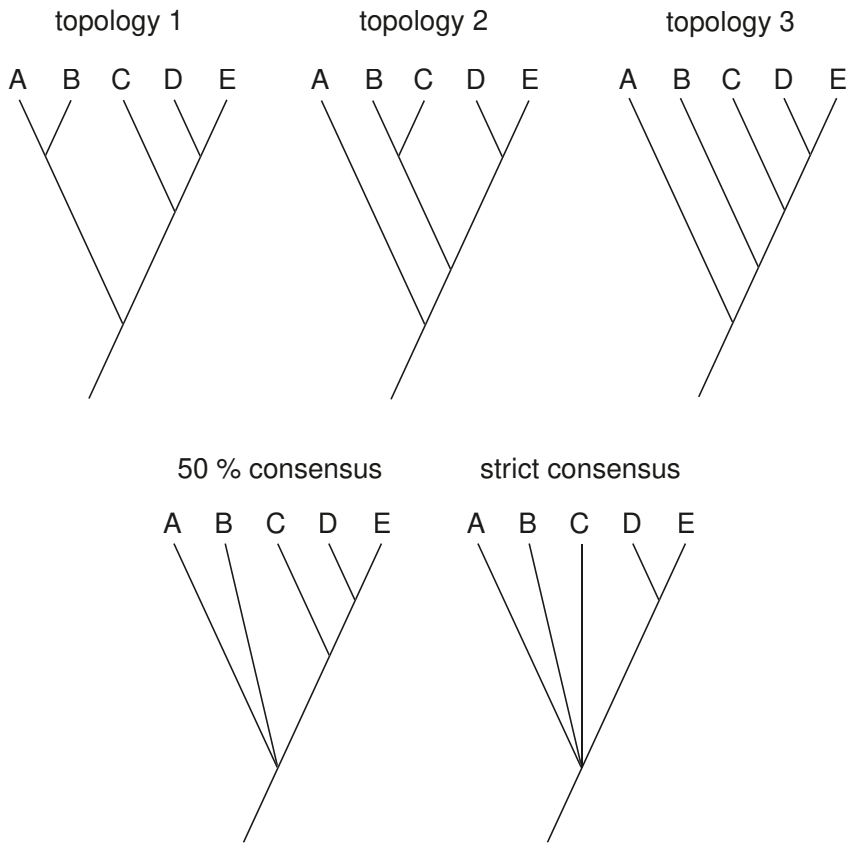


Fig. 58. Three topologies and corresponding consensus diagrams. The only group occurring in all topologies is (D, E). The relation between the taxa A-C remains unresolved in the strict consensus. The group (C (D, E)) occurs in more than 50 % of the cases.

There is also the alternative to retain only such groups (or nodes) which can be found with a given frequency in the set of topologies ("majority-rule-consensus"). If in the example of Fig. 58 groups occurring in more than 50 % of the topologies are to be retained, the group (C, D, E) is depicted as fully resolved monophylum.

With the **Nelson-consensus method** those taxa, which take varying positions in the alternative topologies, are placed in such a way that a compromise is obtained. The Nelson-consensus topology can be found with a clique-method (Page 1989, see ch. 14.5).

With the **Adams-method** those taxa, whose positions vary in the alternative topologies, are attached to the root of the dendrogram, whereas the congruent rest of the topologies is retained. Thus the consensus topology shows above the root only those ramifications which are shared by the different topologies. In the above-mentioned example taxon B could be attached to the root and the group (C, D, E) remains intact. With this method, however, undesirable groupings which are not present in the original topologies can be created. The transfer of taxa with varying position to the root of the consensus topology is not a phylogenetic hypothesis.

Consensus trees have the great advantage to allow also the occurrence of polytomies, which either represent real speciation events or result from a lack of information. It cannot be ruled out that simultaneous multiple speciations occur in nature (ch. 2.5.2). Furthermore, it is to be expected that a set of characters does not contain information for the reconstruction of some of the se-

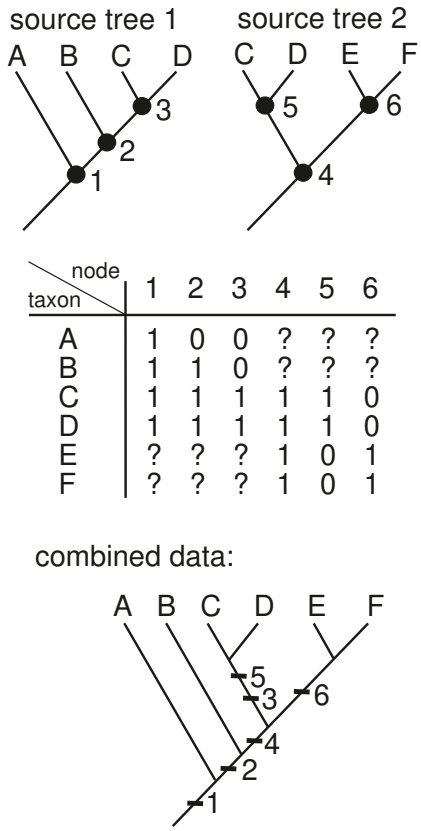


Fig. 59. Example for a supertree construction with matrix representation with parsimony analysis (MRP). Nodes of the source trees are encoded in a data matrix (see text) that is analysed with the maximum parsimony method. The resulting tree (combined data) shows nodes as character states.

ries of speciations and thus the polytomy indicates a lack of resolution due to insufficient quality of the data. As already mentioned (ch. 3.2.2), consensus dendrograms suppress the visualization of incompatible characters and incompatible splits contained in datasets, whereas network diagrams show at least an essential part of the contradictions.

Attention: a prerequisite for the construction of a consensus diagram or a supertree is that different dendrograms have the same probability of being correct. It is a mistake to calculate a consensus from topologies which are based on different datasets with different information content. In this case the consensus topology is less informative than the best of the original topologies. Example: a first topology is calculated from conserved positions of an alignment, a second one from the variable ones or from a second alignment with a more variable sequence. In case the conserved positions are informative and the variable ones are too noisy and saturated with substitutions, the corresponding tree for the latter dataset will contain random groups of taxa supported by noise, impairing the value of the consensus topology.

Note that in general polytomies can have two causes (Wenzel 2002): either data are lacking and therefore a node is not resolved ("**soft**" **polytomies**) or there is contradicting evidence ("**hard**" **polytomies**).

3.3.1 Supertrees and "democratic voting"

Supertrees are a special case of consensus trees obtained from different datasets. The aim of supertree construction is to combine the results of different analyses (different character sets, different species sets, different phylogeny inference methods). A prerequisite is some overlap in the set of species, otherwise grafting of different tree fragments is impossible.

One of the methods that is easy to understand is the matrix representation with parsimony analysis (MRP, Baum 1992, see also Bininda-Emonds & Sanderson 2001). To encode the topologies of different trees, nodes in each topology are numbered (Fig. 59). Each number is used as a character of a data matrix. Then, taxa that are derived from a given node are scored as 1, and the other taxa of the same topology are scored as 0. All other taxa are scored as "?". The matrix is then analysed by maximum parsimony. Characters (nodes) can also be weighted in proportion to the evidence supporting single nodes. Supertree accuracy decreases as the source trees become larger and as taxon overlap between source trees decreases.

"Democratic voting" is an analysis based on supertree methods (or on other consensus techniques) which should be avoided. It is possible to construct a supertree taking results that were published by different scientists or that were obtained with different methods of data analysis. The resulting supertree would reflect some com-

mon features of all available topologies. However, this will in many cases not be the best tree but simply the topology supported by a majority. The supertree topology is *independent of the quality of data* and of the way data were analysed. For example, if several laboratories analyse the same dataset and get different results, the consensus only reflects the portion of shared opinions but not the clades that have the best support by data or that were recovered with the best method. A better way to analyse different datasets is to weigh characters in proportion to the quality of the available evidence and to analyse a combined data matrix (total evidence approach).

3.4 Number of elements of a dendrogram and number of topologies

Number of edges and nodes

When n is the number of terminal taxa ($n \geq 3$), then a dichotomous unrooted dendrogram has

$2n-3$ edges (terminal and internal edges or branches),

$n-2$ internal nodes, and

$n-3$ internal edges.

Therefore: is the number of internal edges (branches) found in a dataset larger than n−3, as expected when analogies or convergences occur, the corresponding splits cannot all be depicted in a dichotomous diagram. In this case it is advisable to use network diagrams (see ch. 3.2.2, 6.4, 14.4) or to apply some criteria for the selection of the best edges.

Number of splits in a dataset

Each possible grouping of species of a dataset produces a split. A **split** is the division of all terminal units of a dataset (taxa, species, individuals) into two groups, usually defined as ingroup (a potential monophylum) and outgroup (the rest of the terminal units; see ch. 3.1). Since always two groups belong to one split, there are less splits than groups. The maximal number of splits which can occur, including those separating only a terminal unit, is:

$$\mathbf{2^{n-1}-1}$$

Taking into account the split between the group of species considered and the rest of the world (a split which is not a bipartition within the analysed set of species), the number is 2^{n-1}.

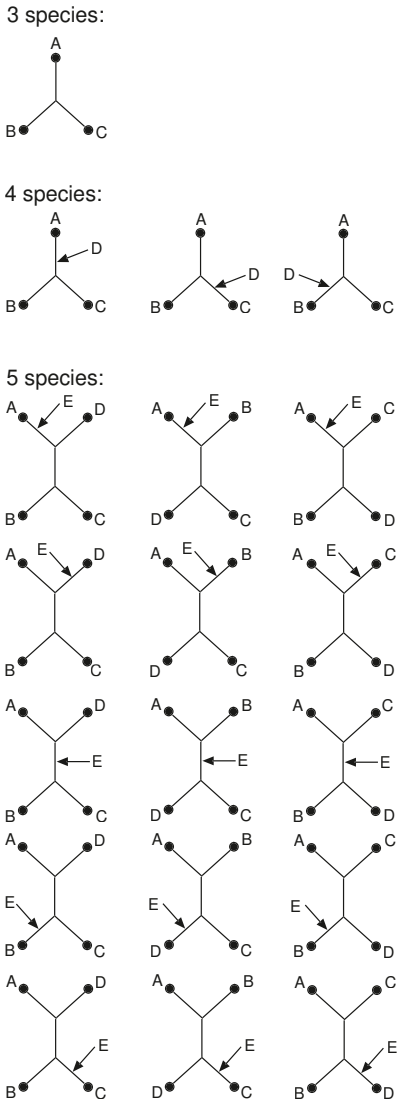


Fig. 60. Construction of unrooted dichotomous topologies from a given number of taxa.

Maximal number of alternative dichotomous, unrooted topologies

If n taxa are given, for $n=3$ only one topology with three edges can be constructed (see Fig. 60). A further taxon can be attached to each of the three edges, and thus with 4 taxa three topologies can be constructed, each with 5 edges. The next taxon can be added to $3 \cdot 5$ edges.

For n taxa we get (Felsenstein 1978a) the number of possible dichotomous and **unrooted** topologies with $B_{(n)} = 1 \cdot 3 \cdot 5 \ldots\ldots \cdot (2n\text{-}5)$, or:

$$B_{(n)} = \prod_{i=3}^{n} (2i - 5) \qquad (n \geq 3)$$

The number of possible dichotomous and **rooted** topologies is:

$$B_{(n)} = \prod_{i=2}^{n} (2i - 3) \qquad (n \geq 2)$$

It follows that with a growing number of taxa the number of alternative topologies which have to be considered for the MP-method (see ch. 6.1.2) increases rapidly. Many computers reach the limits of their capacity when executing an exact search for the most parsimonious topology considering more than 15 species:

n	$B_{(n)}$
4	3
5	15
6	105
7	945
10	$\sim 2 \times 10^6$
15	$\sim 8 \times 10^{12}$
20	$\sim 2.2 \times 10^{20}$

Fig. 61. Table of the number of possible dichotomous unrooted trees with n species.

In mathematics, the search for the most parsimonious tree is also known as the "Steiner problem".

3.5 The taxon

Groups of organisms of nature that are distinguished and named are "taxa". They have to be defined, because the assignment of proper names (e.g., "Mammalia") is not self-evident (ch. 3.1).

In phylogenetic systematics only monophyla are named. Shall a taxon name be accepted by the scientific community and shall it be used in a phylogenetic system, it must be assumed that a corresponding monophylum has been identified (ch. 4.4). Therefore, a taxon should also be a hypothesis of monophyly. If a mentally constructed group (like a taxon) represents a hypothesis that is lawfully derived from empirical observations, it is a special class of mental objects, a "natural kind" (Mahner & Bunge 1997). However, a taxon is not a material object of nature.

Taxa bear proper names. Remember: proper names are not predicators (ch. 1.2). They refer to conceptional or material individuals and do not per se point to properties that are important for the classification of organisms, even though often names of organisms are selected in such a way that they allude to properties: the name *Lumbricus terrestris* names the terrestrial worm; it could be a terrestrial flatworm (terrestrial Tricladida), a soil nematode (Nematoda) or any terrestrial earthworm. However, the name used by scientists refers to only one particular species of Oligochaeta.

Taxa do not exist in nature, and the only phenomenon that can be used to define taxa objectively is common descent. Therefore, a taxon should be a named species or a group of species comprising a common ancestor and all of its descendants (= a clade). Since clades are hypotheses-dependent units, taxa represent hypotheses on descent.

As the names of most groups of animals were introduced into language on the basis of known recent species, many diagnoses of taxa refer to characters of recent species. When subsequently phylogenetically older fossils become known which do not show all of these characters, there is the possibility to adapt the diagnosis of the taxon and to include the fossils, or to exclude them and to name a new, more comprehensive monophylum. This means the traditional taxon name can either correspond to taxa 2 or 3 in Fig. 62. At the moment conventions and tradition decide which alternative will be generally ac-

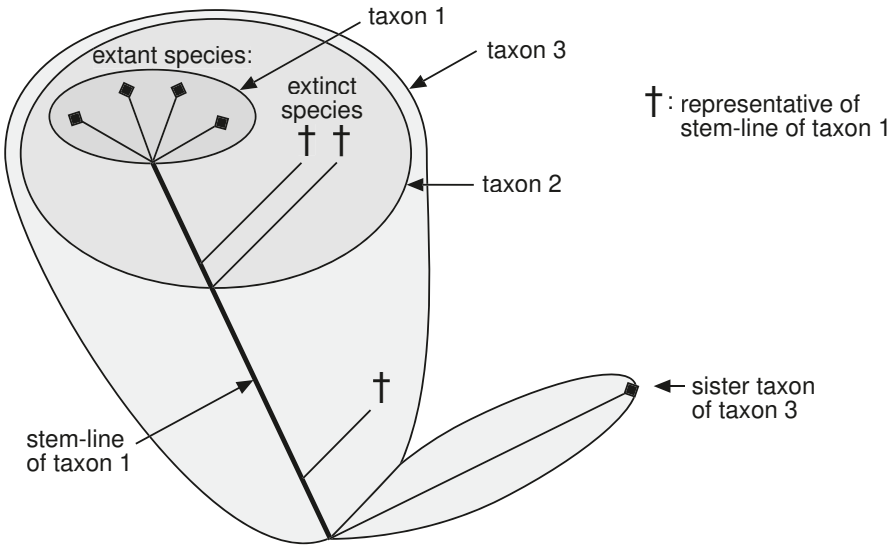


Fig. 62. A taxon has to be defined: where are the boundaries to other taxa? Is the name valid for taxon 1, taxon 2 or taxon 3? The relation of monophyletic taxa to each other are the same as those of twigs and branches of a tree, which can be cut arbitrarily at any place.

cepted. If one decides to incorporate fossils, the **crown group** (= group of recent species including their last common ancestor = taxon 1 in Fig. 62; Jefferies 1979) can be distinguished from the **stem lineage representatives** (see next chapter). Hereby, one always has to name the sister taxon because otherwise the stem lineage would not be limited towards the base of the tree. (The term

Fig. 63. *Archaeopteryx* in comparison with a modern bird. Was the extinct animal a bird or not? This is only a question of the definition of the term "bird". In comparison to recent species the animals baptized with the generic name *Archaeopteryx* had teeth, longer fingers, no large sternum, and a long tail, but also feathers.

"crown group" is ambiguous: one can call a single surviving species of a long stem lineage (like the rhynchocephalian *Sphenodon punctatus* or the Namibian desert plant *Welwitschia mirabilis*) a "crown-group", because it is the recent "crown" of the stem lineage. But the scene of a rich radiation combined with the evolution of successful novelties more closely matches the metaphor "crown").

Usually the most closely related group of *recent* species is defined as the sister taxon (see ch. 2.6). The inevitable consequence is that all fossils of the stem lineage have to belong to the same taxon as the recent species (taxon 3 in Fig. 62). However, a convention for this usage does not exist.

Category: a term specifying the rank of a class (term "class": see ch. 1.2).

Linnéan categories: system of categories used to characterize the rank in the hierarchy of taxa, introduced by Carl von Linné (1707–1778) (see ch. 3.7).

Taxon: a named group of organisms which is distinguished from other groups. Taxa can be species or groups of species. Taxa of a phylogenetic system are monophyletic (term "monophylum": see ch. 2.6).

Crown group: monophyletic group of recent species including their last common ancestor and its descendants, or a species rich monophyletic group that resulted from a successful radiation.

Stem lineage: imagined ancestor-descendant-line leading to the last stem species of a monophylum. The start of the line has to be agreed on by convention.

Stem lineage representative: organism which originated in the stem lineage or that belongs to a stem lineage population and is not considered part of the crown group.

Panmonophylum: a monophylum consisting of the crown group and the stem lineage representatives. Also in this case the start of the stem lineage has to be fixed by convention.

A typical conflict is the following one: it can be argued whether or not *Archaeopteryx* should be included in the taxon Aves. If the Aves are per convention taxon 1 in Fig. 62, a more inclusive taxon would have to be named that corresponds to taxon 2 and includes the fossil. Then by definition *Archaeopteryx* would not be a bird. In this case, the taxon Aves would be defined in phylogenetic systematics with a point in time when the last common ancestor of all recent species existed. This would be the basal limit of the taxon Aves in the tree of vertebrates. On the other hand, if one agrees that all organisms with feathers are named Aves, the boundary of the taxon would not be clear, because it is not known at which point of the stem lineage modern feathers occurred for the first time. Feathered, primarily flightless stem lineage representatives would also have to be called birds (feathered saurians: *Protarchaeopteryx*, *Caudipteryx*, see Swisher et al. 1999). In this case the taxon would include *Archaeopteryx*. It would be defined by an act of classification ("animals with feathers") rather than by systematization.

The naming of taxa obviously requires conventions. These are discussed in ch. 12.

Against this background it is evident that the formulation "taxon **A** is derived from taxon **B**" is logically nonsense (except when used as a metaphor when A and B are species in a dendrogram), because descent is only possible from individual organisms (parents and ancestors in a population), not from taxa or monophyla. Taxa are constructs, which do not live and reproduce as real individual organisms do. And, taxon **A** would not be monophyletic if taxon **B** is excluded. Furthermore, taxon **B** can have side branches (see stem lineage representatives), which do not have to belong to the ancestral population of **A**.

Considering a phylogenetic tree with all recent and extinct species, there is no objective reason to distinguish some groups of species from others calling them **crown groups**. This differentiation is chosen subjectively, in order to distinguish recent from extant species groups or to label a group of species, which originated in a phase of multiple and fast speciations after the evolution of special adaptations ("key characters"). The latter group would be understood to be a "crown" of evolution at a given period in time. Such species-rich groups in the recent fauna are for example the perching birds (Passeriformes, song birds) among the Aves, the sharks (Selachii) among the Chondrichthyes (cartilaginous fishes). Such species-rich monophyla could also be extinct (trilobites, ammonites). Likewise, species-rich groups

living today may become extinct in the future. The concept proposed by Jefferies (1979) implies that one has to name all groups of recent species "crown groups". Then one would also have to accept archaic relict species (e.g., extant species of *Latimeria*, *Nautilus*, *Ginkgo*, horseshoe crabs) as crown groups, the term would be synonymous with "a recent species plus its stem lineage". Furthermore, one gets a whole set of inclusive crown groups (apes, mammals, tetrapods are crown groups), crowns are contained within other ones. It is suggested to avoid the term crown-group.

3.6 The stem lineage

The chronological sequence of ancestors and descendants of individuals or populations (in the sense of reproductive communities), which leads to the last common ancestral population of a monophylum, is called "stem lineage". Therefore stem lineages have a defined end in relation to a monophylum. However, it causes problems to determine the starting point of a stem lineage. Each stem lineage can be traced back to the first living cell. Speaking of the stem lineage of a monophylum however, the series of ancestors and descendants is meant, in which evolutionary novelties of the monophylum that are absent in other taxa occurred for the first time. In phylogenetic systematics the term "stem lineage" is only used for this part of the ancestor series in relation to a monophylum. Therefore, a statement about the organisms that have to be included in a stem lineage requires the reference to a sister group. Since fossils as well as monophyla with recent organisms can be named as sister groups, the selected sister group always has to be mentioned explicitly to make it clear which section of the lineage of ancestors is concerned.

Species or members of species which do not belong to the terminal monophylum and are thought to be located on side branches of the stem lineage or on the stem lineage itself of a reconstructed phylogenetic tree, are called **stem lineage representatives**. The term has proved its worth, because it is not feasible to introduce new names for supraspecific taxa for each speciation along the stem lineage that has been documented with fossils (*supraspecific* are taxa which are composed of more than one species or which are above the category species in the hierarchy of Linnéan categories (ch. 3.7)). The term "stem group" should not be used (Ax 1987, 1988), because it can evoke the association that we are dealing with a monophylum. "Stem groups" are at best groups of stem lineage representatives and they are always paraphyletic.

The restriction of the usage of the word "stem lineage representative" for extinct species, which are direct ancestors of recent species, would result in a near abandonment of the use of the word, because it is highly improbable that exactly such a 'direct' ancestor has been conserved as a fossil. Even if it would really be such an ancestral species it could not be proven, because a closely related representative of a side branch, whose autapomorphies are not conserved or not visible, could not be distinguished from a direct ancestor. Therefore it proved to be convenient to call not only direct ancestors "stem lineage representatives", but also those fossil species which belong to a side-line, even when these species show autapomorphies (Ax 1988). For the species which are direct ancestors, we have the term "**stem species**". A corresponding fossil would be an individual of "one of the stem species".

Stem lineage representatives do not necessarily have to be fossils: in case a living animal would be discovered that proves to be derived from the stem lineage of the Tracheata, for example, a marine animal primarily without tracheae but already with specific characters of the anatomy and with characters of the mouthparts which are considered to be apomorphies of the Tracheata, then the animal would be, with reference to the monophylum Tracheata, a stem lineage representative. This example clarifies that the naming of a specific stem lineage depends on the actual state of knowledge and on conventions: the taxon Tracheata could be redefined so to include the marine species, or a new monophylum comprising the Tracheata and the new species would have to be named (this is recommended to avoid

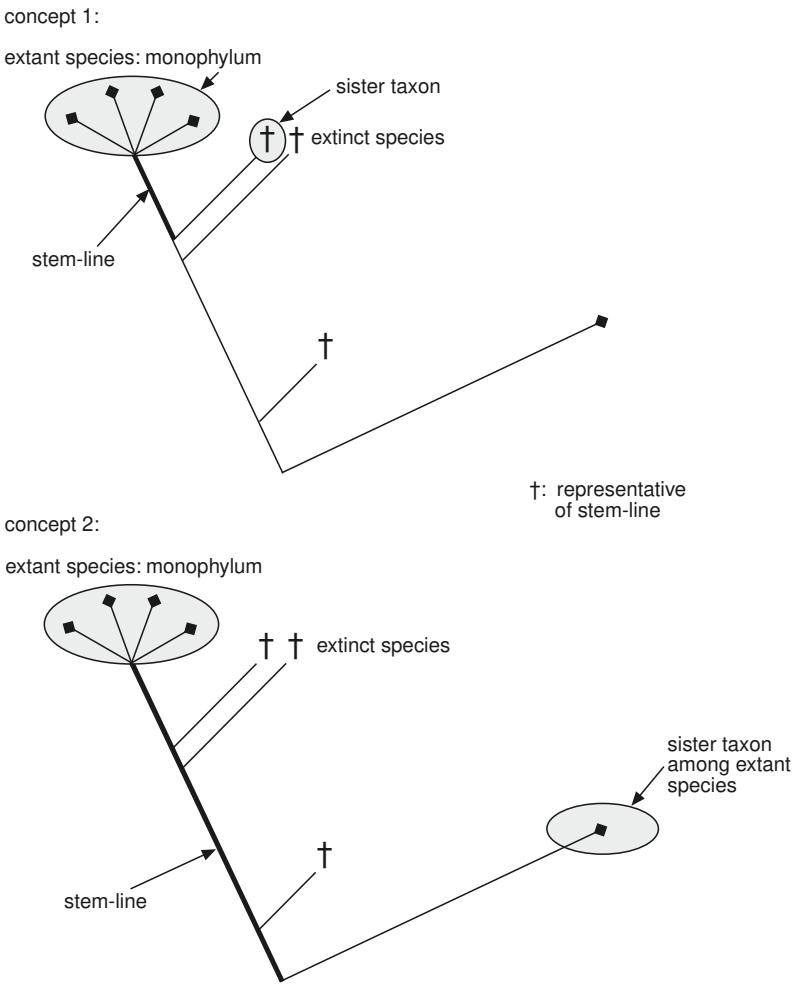


Fig. 64. Example for the limitation of stem lineages in naming a sister taxon.

misunderstandings). In the latter case the concept of a stem lineage of the Tracheata would be reduced to the distance between the new marine species and the terrestrial Tracheata.

Since initially (and often up to now) no fossil stem lineage representatives were found for many "crown-groups", the proper names of the taxa have been applied for the groups of recent species (taxon 1 in Fig. 62). Often traditional scientific names for monophyla that include the stem lineage representatives do not exist. In these cases there is the possibility to name a **panmonophylum** corresponding to a "crown group", which also includes the stem lineage representatives (Lauterbach 1989). The prefix "Pan-" is attached to the name of the "crown group" to designate the more comprehensive taxon. In this way an inflation of taxon names can be avoided. Bear in mind that a sister taxon has to be named also for a panmonophylum in order to determine the basal boundary of the stem lineage.

Example: With the discovery in 1985 of phosphatized Cambrian arthropods, stem lineage representatives of the Mandibulata became first known. These were originally interpreted to be Crustacea (Müller and Walossek 1985). In order to classify these species, the taxon that includes the stem lineage of the Mandibulata (Fig. 65) can be named "Panmandibulata".

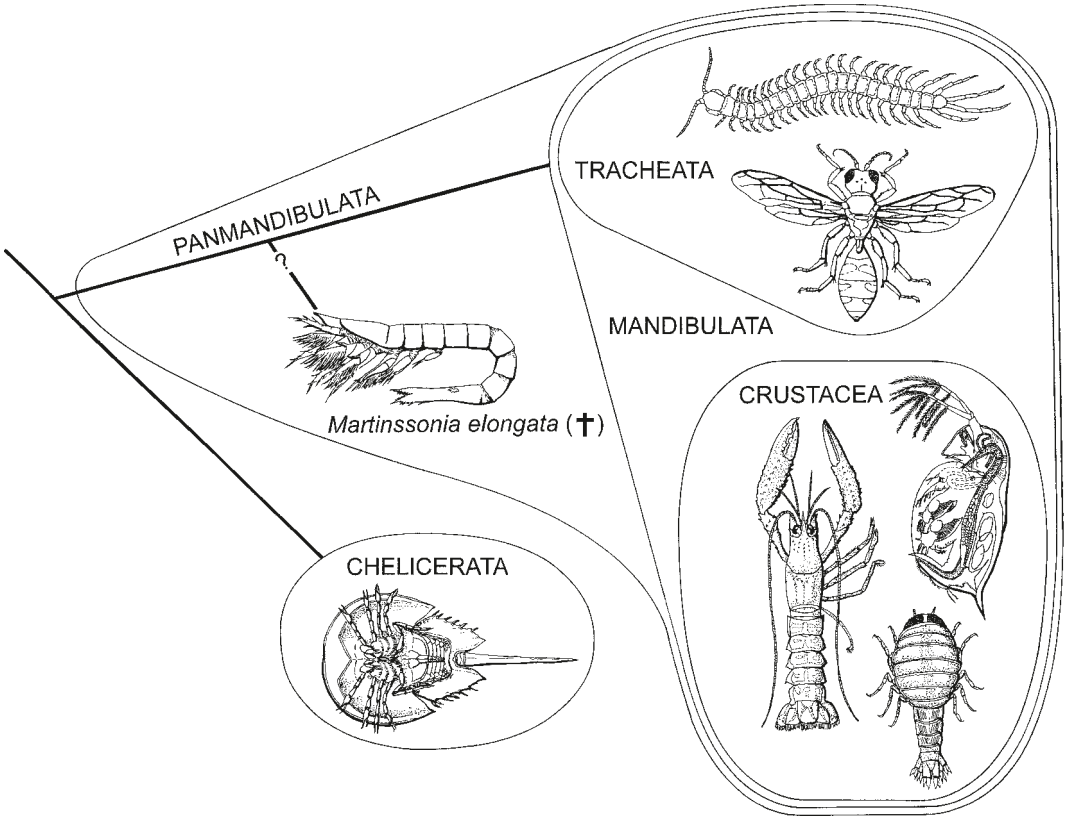


Fig. 65. As long as the evolution of the stem lineage of the Mandibulata is not explored satisfactorily, the taxon including the stem lineage representatives of the Mandibulata is provisionally called "Panmandibulata", because a traditional name is lacking.

Note: the terminal monophylum in Fig. 64 belongs to larger monophyla which include the indicated stem lineages. Therefore these monophyla are not identical. Note also that by selecting a sister taxon in the recent fauna, the actual adelphotaxa are not those which are shaded in Fig. 64, but the two larger (superordinated) monophyla, which include the complete left or right stem lineage, respectively (see Fig. 62).

3.7 Linnéan categories

Carl von Linné (also known as Carolus Linnaeus; 1707–1778) was a Swedish physician and naturalist, who became known in the first place for his work on the classification of organisms in connection with his activities as director of the botanical gardens in Uppsala. He introduced binary nomenclature and used categories to describe hierarchies and therefore can be called the "Father of Taxonomy".

A category is the naming of the rank of a hierarchical level (see ch. 1.2). In traditional systematics, a monophylum is associated not only with a proper name (= taxon name), but also with a category. Categories are thought to have the property to indicate a rank or hierarchical level. In fact, reading about a new, hitherto unknown taxon, specialists can recognize with the help of the category selected by the author(s) of this taxon

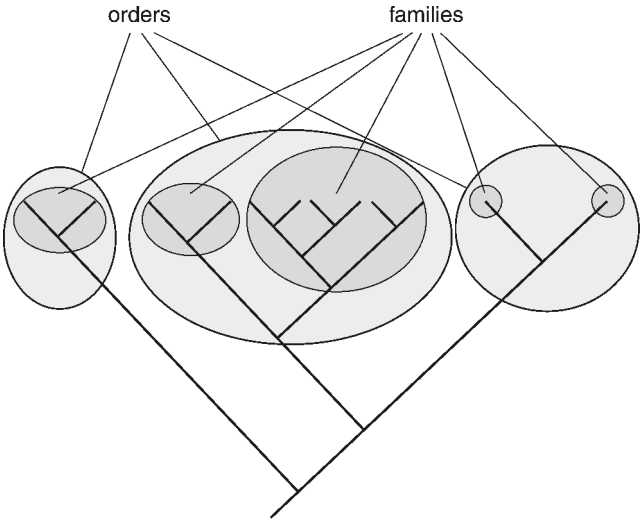


Fig. 66. Categories contain neither information on the genealogical order nor on the extent of taxa.

which assignments into the system of known organisms are excluded, even if the exact placement is not known: a "new family" can only belong to a superordinated taxon of the rank "superfamily" or "suborder", but not to an already described family or to a taxon of lower rank. From this point of view the specification of a category is informative.

It is, however, a mistake to assume that categories allow statements on

- the genetic distance between taxa (see also Johns & Avise 1998),
- the age of taxa,
- the number of included organisms (extent of the taxon),
- the phylogenetic relationships to other taxa.

Example: a comparison of nearly 100 genes of humans, chimpanzees and other monkeys has shown that sequences of humans and chimpanzees are identical in more than 98 % of coding sequence positions. The authors (Wildman et al. 2003) conclude that they have discovered evidence for the inclusion of chimpanzees in the genus *Homo*. This statement has no rational foundation: the assignment of Linnéan categories is artificial and based on *ad hoc* decisions and tradition, and it is therefore not possible to define categories with genetic distance values (see also chapters 3.5 and 12).

The selection of categories is a subjective decision not based on rational logical arguments. Ignoring tradition and using rational considerations you will note that the use of categories really is superfluous and that the desired information can be presented in a better (more precise) way in form of a dendrogram. In any case biologists must know or learn which group of organisms is meant with a taxon name. Nowadays, memorizing categories of higher order together with associated taxon names can be abandoned without consequences for the author, the categories are "taxonomic ballast". For the replacement of lower categories (e.g., "genus" and "family") there are no conventions yet, editors of taxonomic journals usually ask for these specifications. For larger groups, just the name of the group and a dendrogram showing its composition and placement are sufficient.

Due to the traditional use of Linnéan categories, often taxon names are proposed that are empty and redundant. This serves only the preservation of the cascade of categories of higher rank than the species-taxon. An example is the systematization of the species *Symbion pandora* (Funch & Kristensen 1995), discovered in 1995. All supraspecific taxa of this example are "monotypic" , they refer to one species only (Fig. 67).

It can be argued that these empty taxa are wildcards for undiscovered or unknown extinct or-

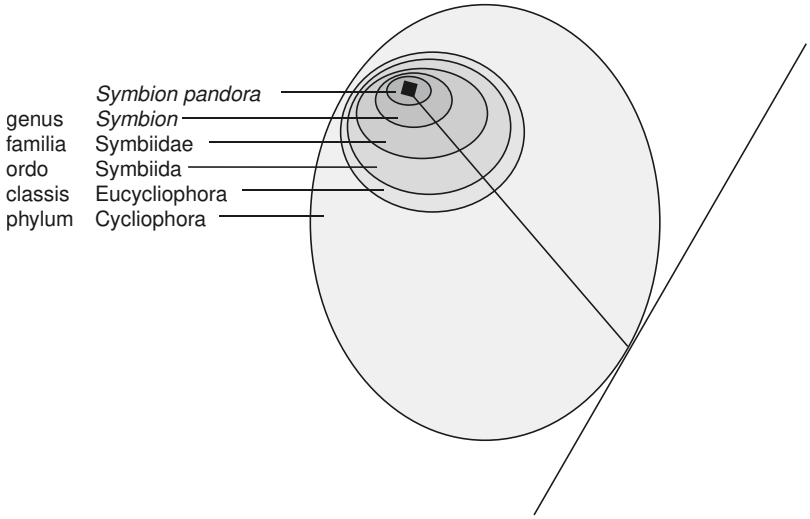


Fig. 67. Redundant taxa erected for the species *Symbion pandora* (categories and names of taxa).

ganisms. There is, however, no empirically founded motivation for the introduction of empty taxa besides the tradition of taxonomy.

A common mistake is the equalization of taxa belonging to the same category. Some ecologists, for example, pretend (to reduce the work load, for convenience, mostly due to the lack of expertise) to be able to determine the diversity of organisms of a landscape (its biodiversity) even if they do not identify species but only genera or even higher ranking taxa. In the same way paleontologists attempt to describe changes of diversity at the level of higher taxa (Fig. 68). In these approaches it is usually ignored that the result of an analysis based on higher taxa depends on two parameters: (a) the more taxa recognized the greater is undoubtedly the biodiversity, but (b) the

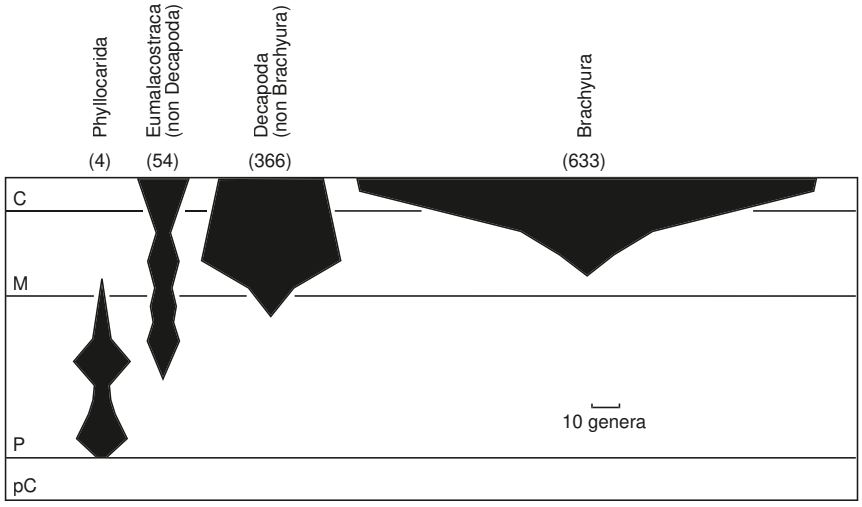


Fig. 68. Example for the description of diversity at the generic level. The fact that the number of species included in a "genus" depends on traditions and subjective opinions and less on the genetic diversity of groups of species is overlooked by many authors. The graph shows the number of described fossil and recent genera of Malacostraca (Crustacea; modified after Moore 1969). pC: Precambrian, P: Paleozoic, M: Mesozoic, C: Cenozoic.

size of the taxa also depends on the classificatory traditions of specialists. Ornithologists combine a far lower number of species to genera and families than entomologists. Therefore, the comparison of diversity on this basis is misleading.

The following table shows the relationships between the terms taxon and category and helps to differentiate them (adapted from Ax 1988):

material objects of nature or groups of objects	**ontological status**	**proper name of the object or of the taxon**	**category**
individual riding horse	material object	e.g., "Fury"	–
a natural herd of horses	material system (reproductive community)	–	–
all existing horses	independent material objects, together a mental "natural kind"	–	–
group of all descendants of the first horse (descendants of the ancestor of our domestic horse)	mental construct	*Equus ferus*	species
group of all recent and fossil species, which are especially similar to the domestic horse (donkeys, zebras, horses) and share a last common ancestor	mental construct ("natural kind")	*Equus*	genus
all horse-like animals, including archaic ancestors of horses (e.g., species of the genera *Orohippus*, *Epihippus*, *Miohippus*, etc.)	mental construct ("natural kind")	Equidae	family
all odd toed ungulates (for example tapirs, rhinos, extinct species)	mental construct ("natural kind")	Perissodactyla	order

Further remarks on the relevance of categories are found in chapter 12.3 (Formal classification).

4. The search for evidence of monophyly

An intersubjectively testable and lasting classification of organisms has to refer to speciation events. Named groups of organisms should be monophyletic (ch. 2.6). In order to identify these monophyla we need information on historical processes, namely on those processes that led to the irreversible divergence of populations (also called speciation processes). Therefore we have to ask which kind of information we are looking for.

4.1 What is information in systematics?

The most important statement of chapter 1.3.5 says that information is a trace left by a process or thing, a trace which is readable by a specific receiver. In systematics, the processes of interest are the "speciation events" (see ch. 2.5). The trace left behind by these are genetic differences between organisms, which can be analysed either directly at the level of nucleic acids ("in the genome") or indirectly with the visible modification of morphological structures ("in the phenotype"). We have to formulate more precisely and state that the informative trace left by evolution in populations of a stem lineage are the apomorphies which can be detected in descendants of the last common ancestral population (see Fig. 69, 76). The receiver of this information is the trained phylogeneticist, the correctly identified apomorphies are the "**phylogenetic signal**" (term: see ch. 1.3.5; see Fig. 5, 69).

The **only information** which exists and can be used are (1) the **apomorphies** present in recent

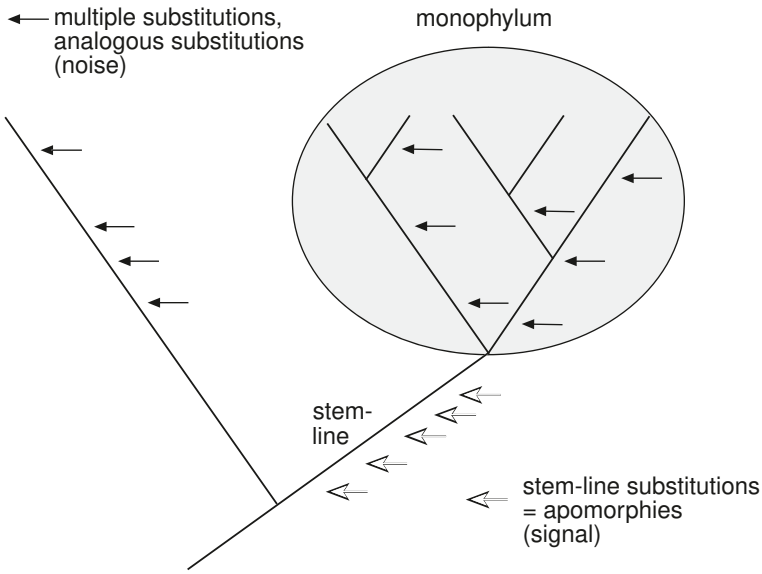


Fig. 69. Evolution produces "phylogenetic signal" and "phylogenetic noise". These traces of phylogeny are only readable for trained persons. The signal consists of substitutions which originated in populations represented in this graph by stem lineages. The signal gets noisy due to the occurrence of analogies, and through erosion, which is caused by superposition by secondary substitutions or novelties that evolved later.

organisms, fossils, genomes of organisms, etc., and (2) similarities that indicate some properties of evolutionary processes. The most difficult task of phylogeneticists consists in the reliable identification of these characters and their distinction from chance similarities. In chapter 5.1 it is explained that this certainty is obtained by estimation of relative probabilities of homology. The more single identities are present, which can be considered to be putative apomorphies due to their co-occurrence in a limited group of organisms, the more information have these patterns of identities (characters). This implies that more informative characters carry more traces of evolutionary events of the stem lineage of a monophylum than less informative characters. In DNA-sequences, patterns which originated from substitutions also tell a lot about molecular evolutionary processes (ch. 2.7.2). The reconstruction of these phylogenetic processes can only be successful when identities mainly consist of apomorphies and not of patterns that are similar by chance.

Evolution does not only produce evolutionary novelties which in the eyes of phylogeneticists are apomorphies, but also novelties which are similar to characters of other, not closely related organisms or which seem to be identical (especially at the molecular level) with characters of other groups. These analogies or convergences are from the phylogenetic point of view "**noise**", meaning modified or false signals. If this fact is not realized by the phylogeneticist, he can mistake analogies or convergences for apomorphies and probably will not be able to infer phylogeny correctly. "Noise" also arises when apomorphies became unrecognisable because further novelties that evolved later changed the original character (Figs. 69, 73, 104). When the apomorphy is completely substituted by later novelties it is objectively not present any more: "the signal is noisy beyond recognition" or "completely eroded". Example: the Amniota are vertebrates with the evolutionary novelty to be able to lay large yolk-rich and hard-shelled eggs, which can develop outside aquatic habitats skipping larval stages. This property is known from lizards and birds. Most mammals, however, produce eggs with little yolk and without shells. The Monotremata prove that also mammals were originally able to lay lizard-like eggs. The characters "yolk" and "eggshell" were reduced secondarily when viviparity evolved within the clade Mammalia. These apomorphies of amniotes "eroded" in the stemline of the Theria. Undoubtedly there also exist corresponding mutations in the genome of the affected organisms.

On this basis it is possible to evaluate the information content of characters qualitatively. **Complex morphological characters** have the following advantages compared to single gene sequences:

- they are the product of complex gene expressions, and therefore visualize differences in many genes and are highly informative,
- data can be acquired without high expenditures for a large number of organisms, which reduces the danger of mistakes due to inadequate species sampling,
- they allow the analysis of the adaptive value of novelties, which is significant for considerations of the plausibility of hypotheses (not for the substantiation of sistergroup relationships; see ch. 10).

But they also have disadvantages:

- they are present in limited numbers,
- without very costly and time-consuming genetic analyses it is usually not possible to correlate morphological differences with the corresponding number of substitution events and thus with the true genetic divergence. Differences and novelties cannot be recorded quantitatively with a universal unit of measurement.
- Since constant characters are usually not recorded and there exist no methods to define a set of comparable characters it is difficult to analyse objectively differences in character variability.

However, the importance of the high information content of complex characters prevails. In the "pre-molecular" time of biology, these were the basis for the identification of most larger monophyla which are still accepted today. This fact should caution all those who only "believe" in molecules.

Sequences have other advantages:

- they allow the differentiation of cryptic species which are morphologically identical,

- they enable the discovery of homologies for species which do not show any morphological similarities (compare for example fungi with animals),
- neutral sequences furnish data on relationships which may be independent of selection pressure,
- they evolve more constantly (less desultory) because many mutations are exposed to little or no selection pressure,
- differences can be quantified exactly and can be traced back to single historical substitution events as long as divergence times are small.
- There are many more characters (sequence positions) available than comparative morphology can yield. However, it must not be overlooked that these single characters are poor in complexity (when the character is a sequence position) and thus have little information content.
- The process of sequence evolution can be modelled (see ch. 14.1). However, it is very difficult to find out whether these models are realistic.

The disadvantages of sequences are

- the limited information content of the individual nucleotides (there are only 4 alternative characters), which can only be compensated using long sequences,
- the lack of a correlation with evolutionary adaptations to the environment (in most cases),
- problems with the alignment of variable sequence regions (ch. 5.2.2.1),
- the genealogy of individual sequences does not have to correspond to the phylogeny of the majority of the genome,
- there is the danger that contaminations are sequenced. As long as there exist no data on related species in gene banks, the origin of a sequence cannot be checked.

One must not forget that environmental factors act through interactions with the phenotype. The adaptive value of most novelties has to be analysed at the level of morphology, physiology or behaviour. For the reconstruction of evolution we need knowledge on the morphology and mode of life of the animals, for the reconstruction of phylogeny sequences may be sufficient.

4.2 Classes of characters

4.2.1 Similarities

In ch. 1.3.7 it has been explained that a "character" is not a "fact" but a mental construct, a hypothesis for perceived similarities. Four classes of similarities can be distinguished:

a) **superficial similarity,** which is based on inaccurate observation. On closer view it can be seen that the structures are assembled by totally different components. An initial statement of homology may become after more detailed analyses a statement of convergence or analogy.

b) The similarity is also present in some details and can be determined intersubjectively. It can be traced back to processes which occurred independently from each other in different ancestors of the compared organisms (**chance similarity**, **analogy**, e.g., the occurrence of black spots in the fur of domestic animals, the appearance of point mutations at the same sequence position). Stalked eyes (Fig. 72 top) can have different functions: they can serve the improvement of vision or (in the case of some Platystomatidae) increase the attractiveness of males. Therefore different selection factors may have a similar effect. Many similarities however, are non-random adaptations to the same environmental factors.

c) When the processes that shaped phenotypes were influenced by the same environmental factors, a similarity in unrelated organisms is a **convergence**. This can also be regarded as the result of chance, because suitable organisms as well as selection factors with similar effects are not present everywhere and at all times. The fact, however, that comparable selection factors have similar effects on non-related organisms (when these organisms are sensitive for the same factors) is not pure chance, because adaptations occur according to the same physical or biochemical laws

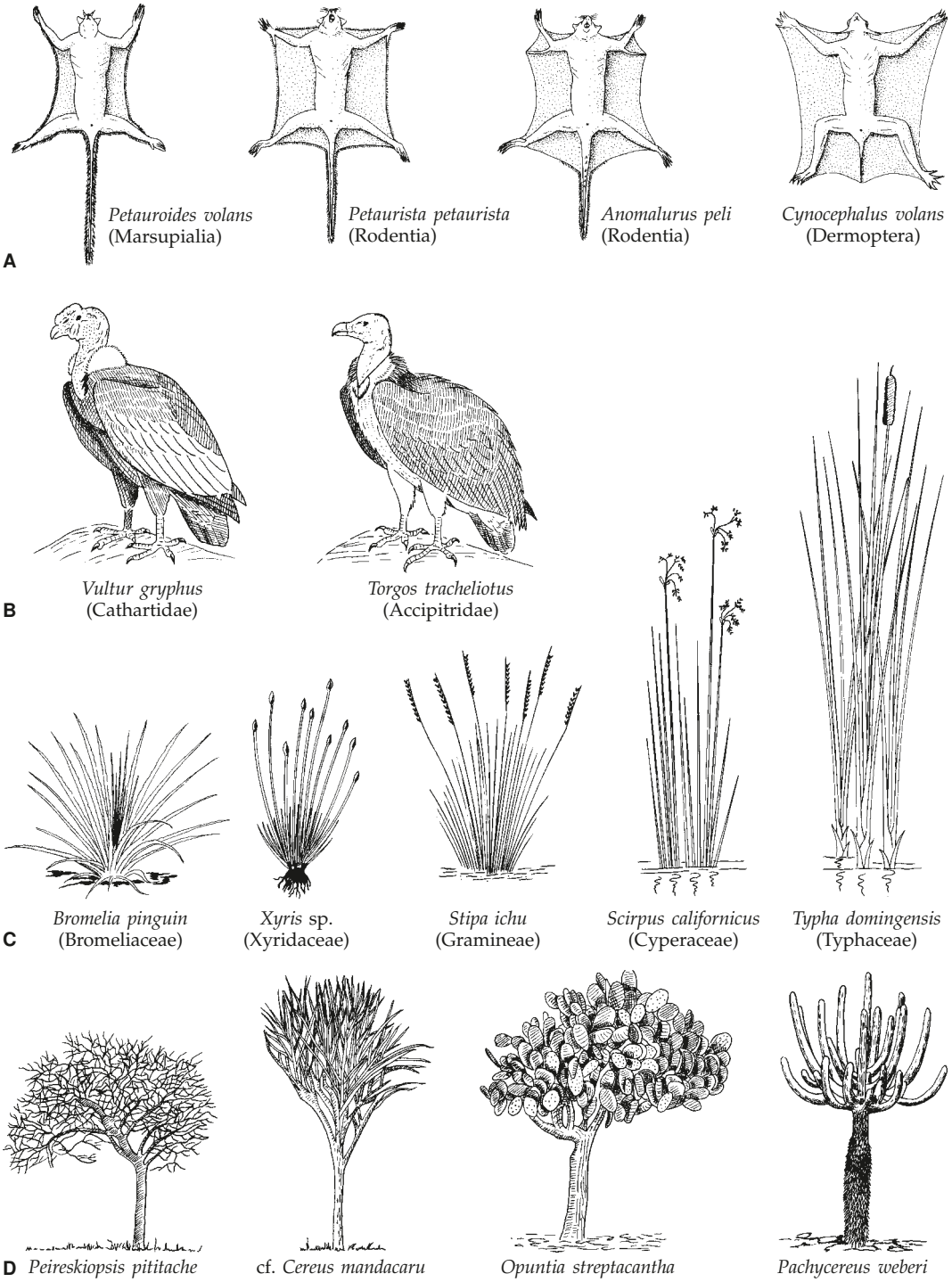


Fig. 70. Examples for convergences. **A.** Jumping and gliding mammals. **B.** Vultures of America (Cathartidae) and vultures of Eurasia and Africa (Accipitridae). **C.** Plants with grass-like growth form. **D.** Tree-like growth form in Cactacea (modified from Koepcke 1971-1973).

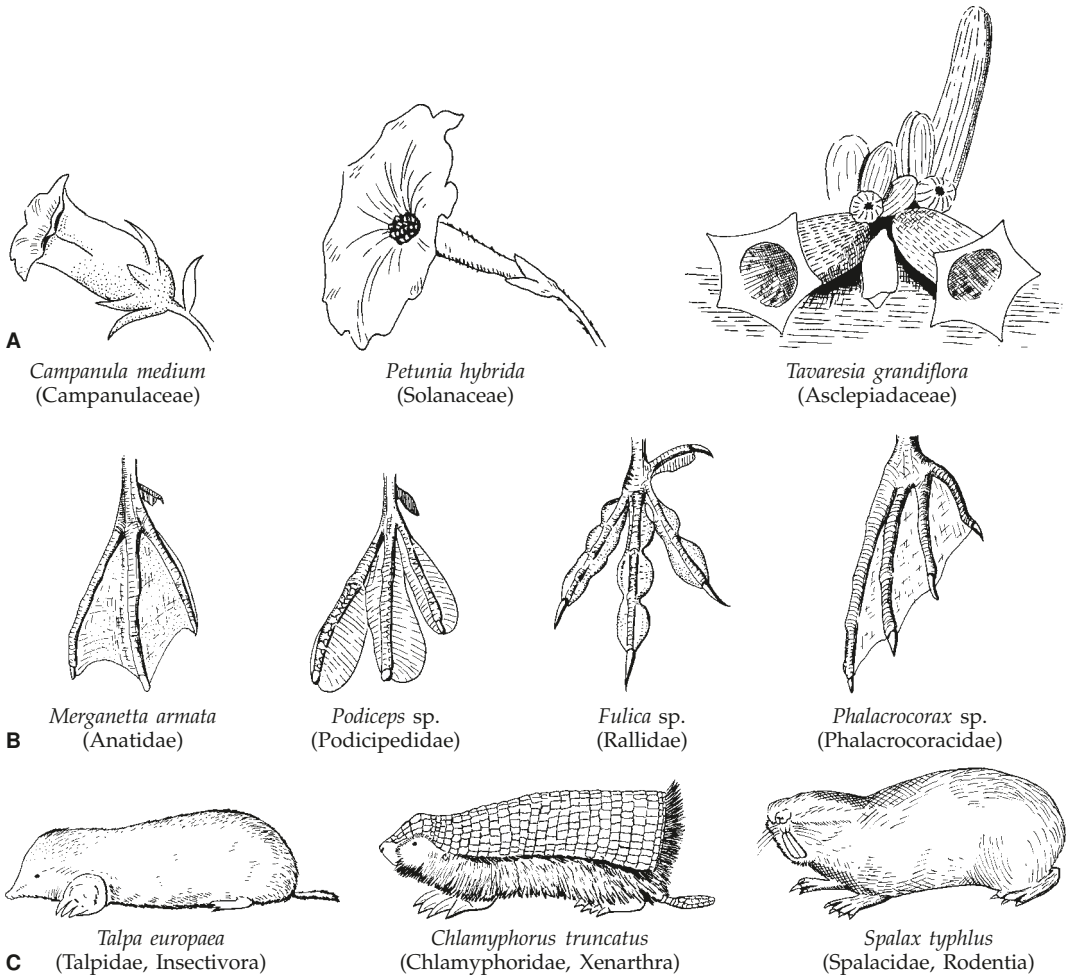


Fig. 71. Examples for convergences. **A.** Tubular flowers growing laterally on plant. **B.** Feet of water birds. **C.** Fossorial mammals (modified from Koepcke 1971-1973).

(e.g., laws of optics, hydrodynamics, physical chemistry). When a convergence evolves in related organisms, often the same anlagen are shaped into similar forms.

Superficially similar characters that are definitely not homologous can be clearly discerned from those that evolved from the same homologous ancestral character. For example, piercing styliform mandibles evolved convergently in mosquitoes and tree bugs from the same appendages: the mouthpart primordia are homologous, the details of shape and function are convergences. Such convergences are also called **homoiologies**.

When species are so closely related that the organisms have nearly the same appearance, one talks of **parallelisms**. This term implies that the evolution of a homologous character took place in parallel in two species, its modification is similar but not homologous. The decision where to draw a line between homoiology and parallelism is subjective. "Convergence" can also be used as the main term that includes the other ones. (The word "parallelism" is also used for other notions, for example, for the parallel evolution of parasites and hosts). Attention: the term "convergence" refers to the evolutionary process as well as to the result of the process, the adapted structure itself.

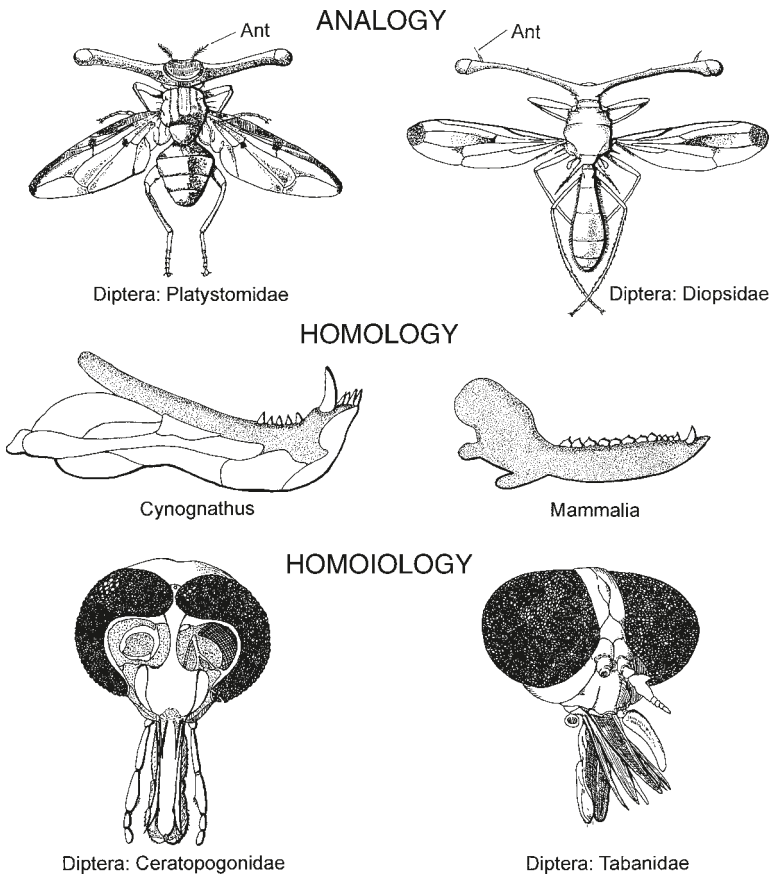


Fig. 72. Examples for analogies, homologies, homoiologies. The fact that the stalked eyes in Diptera evolved convergently can be seen in the different positions of the antennae (Ant). – *Cynognathus* sp. is a fossil from the Triassic, which is classified as representative of the stem lineage of the mammals. The dentale (stippled) is homologous to the lower jaw of modern mammals. – Within Diptera, the mouthparts are homologous structures, which developed convergently several times to similar piercing instruments.

In practice, the term "analogy" is mostly used for chance similarities but also for congruencies caused by selection, whereas the term "convergence" always implies an adaptation to the same environmental factors. As in nature there exists a continuum from "neutral" to "highly effective" selection factors, there are no sharp boundaries between the terms "purely accidental analogy" and "convergence".

Analogies and convergences are non-homologies. The fact that a homology cannot be substantiated with certainty (see ch. 5.1) is not a sufficient reason for calling any similarity a non-homology. The diastema, for example, a similar gap between incisors and molars in the dentition of rabbits, rats, horses and e.g., roe deer is as character so poor in structure (just a "gap") that the homology of the gap alone cannot be proven. The diastema can actually be a homology (in closely related species), but also a convergence (comparing rabbit and horse, for example). The **recognition of non-homologies** (analogies or convergences) can be achieved in different ways:

Case 1: when species 1 and 2 show the similarity X_1 and X_2 (e.g., stalked eyes in Fig. 72 top), it has to be shown that character X_1 (of species 1) is homologous to a character Y of species 2 (that is not X_2) or of a third species (e.g., the stalk of the eye of the Platystomatidae is homologous to an area between the insertion of the antenna and the

eye in Diopsidae or other Diptera), and that character X_2 of species 2 is not homologous to character Y but homologous to Z (requires a proof that the stalk of the Diopsidae in Fig. 72 does not correspond to the area between antenna and eye of other flies, but is homologous to the area between the original insertion of both of the antennae). – **Case 2:** the hypothesis of homology for the similar characters X_1 and X_2 of the species 1 and 2 cannot be brought in accordance with the phylogenetic position of the species and the distribution of plesiomorphic character states, because the species are found on different branches of the phylogenetic tree and the most parsimonious explanation is a parallel origin of X (see characters 2 or 3 in Fig. 78). Example: skin folds used as wings are present within mammals in flying squirrels (*Petaurista*, family Sciuridae), belonging to the Placentalia, and in sugar gliders (*Petaurus*, family Petauridae, Fig. 70). The anatomy of the latter clearly characterizes them as marsupials. – **Case 3:** two structures similar at first view show in a more detailed analysis no congruence which could be homologous and apomorphic for the same putative clade. For example, the similar details occurring in the wings of bats and in the wings of birds are only features of the front limb of all tetrapods, which can also be seen in species which do not fly. There is no evidence for the homology of those details which are adaptations to flight. Wing colouration in the American butterfly *Limenitis archippus* (viceroy butterfly) is very similar to that of the monarch *(Danaus plexippus)*, but the species differ in many morphological details such as wing venation. A more formalized argument: of three species A, B and C, two (A and B) share one similarity, while another pairing (B and C) shares many more common characters, indicating that B and C are closely related while the similarity of A and B must have evolved in parallel.

d) **Homology**: the similarity is not accidental but originated from the same source. Homologies in biology are nearly always congruencies based on the presence of identical or partially mutated copies of the same DNA- (or RNA-) molecules of a common ancestor. Note that two patterns can be dissimilar and nevertheless homologous (ch. 4.3.1).

The differentiation between homology and analogy is often attributed to R. Owen, who defined the terms in a glossary (Owen 1843): "HOMOLOGUE The same organ in different animals under every variety of form und function.". The terms, however, are older (see Panchen 1994). – Definitions, which correspond to the one favoured here, are for example, those of Van Valen (1982: "correspondence caused by continuity of information") or Osche (1973: "homologous are hence structures whose non-random correspondence is based on common information").

On principle, homologies are based on inheritance. In most cases it is DNA which is copied. Morphological structures which are built through the concerted action of genes and gene products are homologous because similarities originate from the presence of copies of the same "blueprint", which may be a complex genetic developmental program. The more is known about the interactions between regulatory genes and structural genes, the more precise is a statement on homology. Sometimes the developmental genes are homologous, but the structural genes that are activated, and vice versa. Homologous developmental cascades may change their function in the course of evolution.

The correspondence between homologous structures does not have to be perfect, the "copies" can diverge in detail from each other. However, as long as it is perceptible that similar structures are probably *copies of the same original*, it can be assumed that they are homologous (see criteria for homology, ch. 5.2.1).

For a specific arthropod appendage, for example, the following homologies may be discerned, each requiring an inherited coding:

- site of the anlage of the leg bud on the trunk,
- point of time for the growth of the leg during ontogeny,
- structure of signal molecules triggering leg growth,
- anlage of branches, of exites and endites,
- number of joints,
- shape of an article,
- cuticular structures,
- anatomy of hair sensilla,
- number of the hair sensilla,
- structure of propioceptors,
- muscle anlagen, insertion sites,
- innervation,
- formation of ectodermal glands,

– place for the anlage of chromatophores,
– characters of cellular organelles,
– and so on …

Many of these details are independent of each other. The characters "fine structure" and "number" of hair sensilla can vary independently from each other in nature; the fusion of two articles does not have to influence the presence of sensilla; the development of endites does not necessarily have consequences for the number of articles, etc. This only indicates that many different genes, which can each be homologized individually in different species, are involved in morphogenesis.

Please note that the definition of the biological homology concept does **not include any statement on the function** of structures (Fig. 101, ch. 5.2.1). Structures can be homologous independently of their use. In humans, males have nipples which apparently have no function, but their anlage is part of the genetically fixed "building program". Their assemblage is obviously determined because the homologous nipples in "females" are of essential importance for the reproductive success and no mechanism evolved to suppress their formation in males. Consequently, an analysis of function alone does not allow conclusions on the probability of homology. Also remember that convergences are mostly composed of non-homologous adaptive features with the *same* function.

Attention! The term "homology" has two meanings: it can (a) name what is really present in nature in form of copies of ancestral DNA-sequences (and their expressed products), or (b) *what we think* is the copy. In practice, we cannot distinguish between (a) and (b), the "homology" is always a hypothesis. – Sometimes authors talk about "monophyletic characters". This expression should be avoided in order not to blur the clearly outlined meaning of the terms "monophyly" and "homology". Only groups of organisms are monophyletic, characters are *homologous*. In many cases the homologies even are not "monophyletic" in the desired sense, because complex characters do not originate from a single ancestral population but evolved stepwise in several consecutive species. For the "monophyly" of homologies we have the precise term "apomorphy".

The systematist has to identify real homologies and to distinguish them from chance similarities. How this differentiation is achieved will be explained in chapter 5.1.

The homology concept presented herein, which is indispensable for systematics, implies that homology is an "all or nothing" concept. A detail is either "a hundred percent" homologous or not homologous at all. A statement of the type "a structure is 50 % homologous", can only have the meaning that 50 % of the components of the structure are not at all, but the remaining elements are "a hundred percent" homologous. – The distinction between a morphological, a biological, and a historical homology concept is irrelevant for the systematist, he rather has to understand the ontology. The "**historical concept**" stresses the phylogenetic origin. This is the phylogenetic cause for the occurrence of homologies. The "**morphological concept**" refers to the structural identity: this is the trace left by evolution. The "**biological concept**" emphasizes the presence of the same developmental constraints which cause the ontological development of a morphological character by autoregulation. The systematist does not necessarily have to know the developmental mechanisms, but he has to evaluate the quality of the characters he is using (ch. 5.1). Also he does not need to know the phylogenetic origin of a character in order to postulate a hypothesis of homology: it is not necessary to know those first fossil animals which had feathers, or to know the correct phylogeny of birds, to identify the feather as a homology occurring in birds.

4.2.2 Classes of homologies

Constitutive and diagnostic characters

Groups of organisms can be recognized by their characters when these are homologies. Two classes of characters which can be used for identification purposes have to be distinguished:

– **constitutive characters**: these are evolutionary novelties (= **apomorphies**) which evolved in a stem lineage of a monophyletic group. They are evidence for monophyly.
– **diagnostic characters**: these are unique characters which can be used for a determination key. Diagnostic characters can, but do not necessarily have to be constitutive at the same time.

Example: among the European amphibians, the urodeles (salamanders and newts) (Urodela, Caudata) can be identified on the basis of their tail

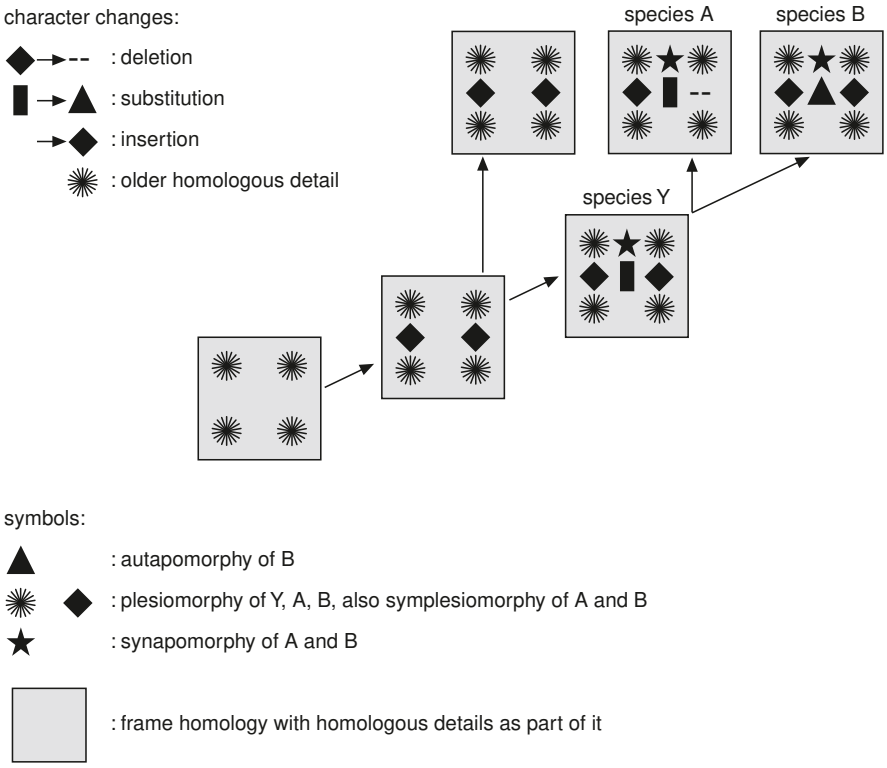


Fig. 73. Possible modifications of details within a frame homology. Attention: in practice the terms "insertion", "deletion" and "substitution" are used to name the *processes* causing the change as well as the *result* of the processes. The meaning can be inferred from the context. The apomorphic character shared by A and B (synapomorphy) is the character still discernible in the recent species. The sistergroup relationship of A and B can be substantiated with this character (see also Fig. 76).

vertebrae, which are lacking in anurans (frogs and toads). This character is diagnostic but not constitutive, because tail vertebrae are an ancient character of Tetrapoda and also occur in other tetrapods not belonging to the Amphibia. This character is a phylogenetically old character, a plesiomorphy (see below). The jumping hind leg of frogs (Anura), the specialized pelvis bones and the strong musculature are at the same time diagnostic and constitutive characters of frogs. These peculiarities evolved only in the stem lineage of frogs.

This distinction is often neglected in taxonomic descriptions, where a "diagnosis" usually is a mixture of constitutive and diagnostic characters. Taxonomists must learn to list separately those characters that are evidence for monophyly of their taxa.

Frame homologies and detail homologies

For complex morphological characters a homology statement means that **only the corresponding (identical) details**, which have really been inherited from a common ancestor, are homologous. Complex structures can also (and usually will) contain details which vary between organisms and are not homologous. The scheme in Fig. 73 illustrates these circumstances.

What is called a "homologous character" by morphologists can be (a) a complex **frame homology** containing novelties as well as older detail homologies, or (b) a single **detail homology**. Since the frame homology can also be part of an even larger organ or organism there exists a hierarchy of encaptic homologies (that is, a homologous detail is part of a more complex homology and this is possibly again part of a larger organ etc.).

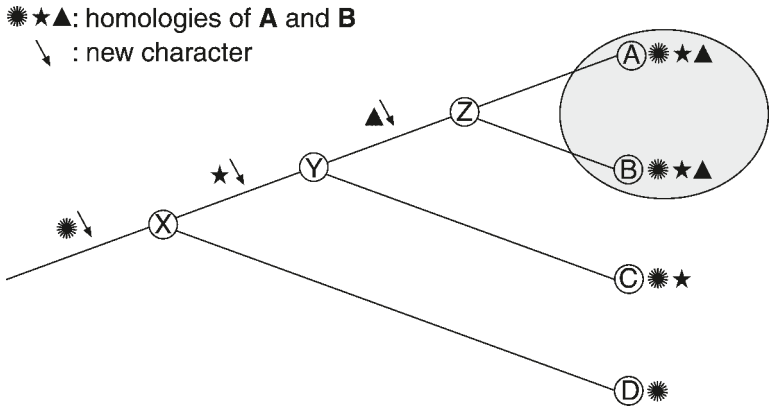


Fig. 74. Different homologies occurring in two species may be of different age.

Example: the complete morphology (including inner organs) of humans and chimpanzees is composed of a large number of detail homologies. After deduction of the peculiarities of the respective species this whole morphology is the largest homologous pattern shared by humans and chimpanzees, thus a highly complex pattern with a hardly countable wealth of details. A subordinate homology belonging to this pattern is the frontal appendage with hand and thumb, which includes, for example, the genes activated for the synthesis of the horny material of the thumbnail. Ignoring their structural complexity or their chemical composition, these structures are only homologous if their ontogenetic assemblage is coded by homologous structural genes (while the genes that activate the construction cascade must not necessarily be the same). Homologous genes have to regulate the formation of a bud for the frontal appendage at a certain position of the body, thereon the anlage of a thumb and later the construction of the corresponding bones and the anlage of a thumbnail at the tip of the thumb.

The distinction between frame homologies and detail homologies can be justified with the phenomenon that genetic developmental programs change during the course of evolution because they are modified by small mutations. When in two organisms details of developmental programs are not all identical or of the same function, or if they activate different subprograms, the developmental programs can nevertheless be copies of an older original and thus would be homologies. We have to discuss separately the homology of single developmental genes, of gene cascades, of structural genes and their products. If this fact is known, erroneous statements on homologies of large organs based on the observation of single gene expressions can be avoided (see examples in ch. 5.2.1). Phrased in an abstract way, homologies can be noisy copies of an original: the complete copies are the frame homologies, but they can contain noisy (modified) details.

Examples: Frame homologies and (in brackets) the relevant details, which may be modified in a non-homologous way: the anterior limb (number and shape of bones of the hand), developmental program for the morphogenesis of the anterior limb (single involved genes), the eukaryotic cell (presence and structure of organelles), the 18S rDNA-gene (nucleotides). See also Fig. 97.

Attention: again we have to distinguish between the real fact and the perception of the fact. When a frame homology is nearly completely modified by subsequent substitutions in comparison with its first state, it may happen that a hypothesis of homology cannot be supported any more with empirical evidence. Although the very different patterns originated through a series of copies from a common original, this fact may not be recognizable any more (see also ch. 4.3.1).

R. Riedl (1975) was the first to draw attention to the hierarchy of homologies in an organism. He distinguished frame homologies from subordinate homologies, the smallest unit of which he called minimum homologies. Riedl also suggested to introduce the sum of minimum homologies as an estimate for the complexity of a frame homology (see ch. 5.1). The impor-

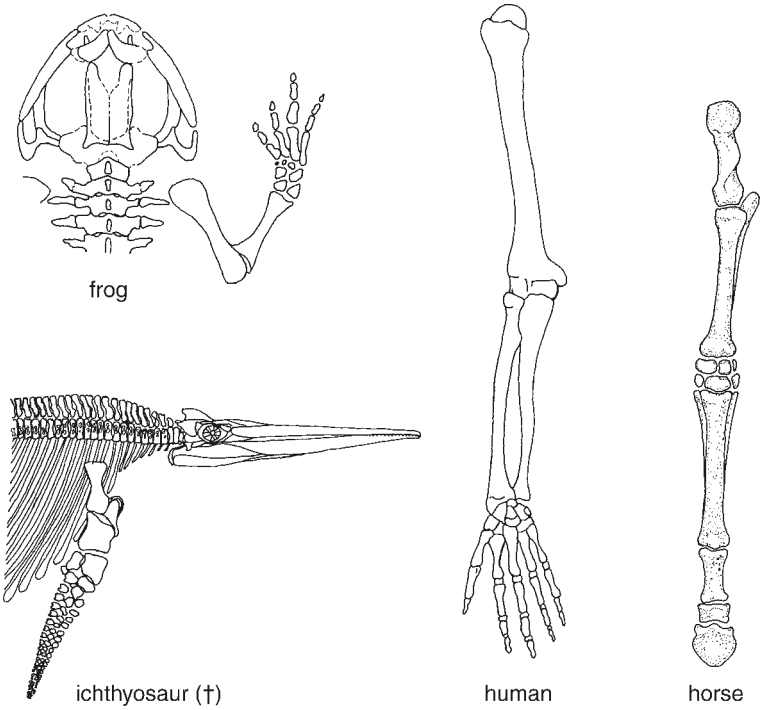


Fig. 75. What is a homology? Anterior limb of frog, human, ichthyosaur, and horse. Is the arm of humans "more homologous" to the arm of the frog or to the one of the horse? Only *identical* genetic material is homologous (see text).

tance of details for the function of frame homologies is discussed by Riedl in context with the term "burden". This means that organs can be functionally dependent of a certain detail and are not free to vary. The consequence are constraints for the evolution of character states. This "burden" cannot be estimated directly for the practice of systematics, for example to describe quantitatively the probabilities for character state changes, but the effect of such constraints becomes visible, because characters under high selection pressure are less variable.

Detail homologies of different age

A homology statement on characters of two organisms implies that these organisms have common ancestors. However, nothing is said about whether the character stems from a distant or a closer ancestor (Fig. 74).

Considering taxon {A,B} in Fig. 74, three homologies, which were inherited from species Z, can be demonstrated for the species A and B. For the species of the superordinate taxon {A,B,C} only two, for {A,B,C,D} only one homology is found. We see that the number of homologous details decreases the more inclusive a taxon is. The same is true for any complex character which is homologized: the frontal leg of humans is very similar to the one of chimpanzees, there is correspondence in many details; therefore the arms are said to be homologous. In the language of biologists it is also said that the arm is homologous to the anterior leg of a horse or of a frog (Fig. 75). This only means that in the frontal limbs of frogs, horses, and primates there exist details which were inherited from a common ancestor. The homology "arm of humans – arm of chimpanzees" is, measured on the basis of the number of shared identical details, much more comprehensive than the homology "arm of humans – pectoral fin". With the extension of the homology concept to other hierarchical levels the composition of the homology changes.

Inherited homologies and other copies

The homology concept described above corresponds to the **inherited homology** and is the exclusive significance of the word "homology" in this book. The homology of technical developments, of literary transmitted information, of learnt behaviour etc. does not arise through copying of nucleic acid sequences and is of no significance for phylogenetics.

Expressed and non-transcribed sequences

The homology concept also includes sequences which are usually not transcribed. These may be non-coding sequences, defective or inactive genes. The existence of inactive genes became known before the analyses of genomes was possible, because sometimes their presence is visible as in the cases of **atavisms** (in mutants showing older character states: three-hoofed horses, humans with elongated vertebral columns, zebra-like stripes in mules). When in the swordless fish species *Xiphophorus xiphidium* (Poeciliidae) of the group of sword-tail tooth-carps the growth of a sword is induced through treatment with hormones, genes are activated experimentally which are homologous to those of other *Xiphophorus*-species. Such genes ("cryptotypes", "latent potentials": Saller 1959, Osche 1965, Sudhaus 1980) can also be reactivated lastingly. Examples: reappearance of the second lower molars in Scandinavian lynx (Kurtén 1963) or of the mandibular palp in hognathiid Valvifera (marine isopod crustaceans) (Poore & Lew Ton 1990).

Homonomy

Homonomy is an "iterative" or serial homology occurring in a single organism. It evolved from duplication of an organ or of a gene. The walking legs of a centipede (Chilopoda) or the copies of an rRNA-gene are homonomous structures. The homonomy of morphological characters is probably based in most cases on the repetitive activation of the same genes at different sites of the body (remember that each cell of the body carries the same genes!). As long as the homology of a duplication cannot be verified, only the single character of which several "copies" may exist (e.g., the general construction of the leg of a centipede) is relevant for phylogenetic analyses. However, when deviating details occur in individual "copies" of complex homonomous structures, these details can again be treated as independent characters that may be homologous in different species. For example, a first maxilliped of a crab (Decapoda: Brachyura) is as a "thoracic leg" homologous to a cheliped or to walking appendages of the same individual. However, special details of the morphology which occur only on first maxillipeds can be homologized with the same details of first maxillipeds in other species of Decapoda.

Apomorphy and plesiomorphy

An evolutionary novelty is an **apomorphy**. A **frame homology** containing an apomorphy (Fig. 73) has an "apomorphic character state" or is a "derived character". A statement on the identification of an apomorphy has to include always a group of organisms in which the novelty occurs for the first time. Older details replaced by the novelty but present in other organisms are **plesiomorphies**. The frame homology of these organisms has a "plesiomorphic character state" or is a "primitive character".

The evolutionary novelty is the result of mutations (insertions, deletions, substitutions, inversions, gene duplications) or of gene transfer. Not all novelties have visible consequences in the phenotype, many mutations are "neutral" (see ch. 2.7.2.2). Unique modifications of morphology, physiology or of behaviour that are based on genetic changes and thus are inheritable are also (and rightly so) called apomorphies, even though the genetic basis of such changes is rarely known.

The term "novelty" implies a relation to time: the naming of an apomorphy requires the reference to a specific level in time, in which a historical stem lineage or stem population with new genetic variants existed.

To avoid the suffix "-morphic" when physiological or ethological character states are discussed, some authors (mainly entomologists) use the terms "**plesiotypic**" and "**apotypic**". However, most scientists are employing Hennig's terms (plesiomorphic, apomorphic) for any type of character.

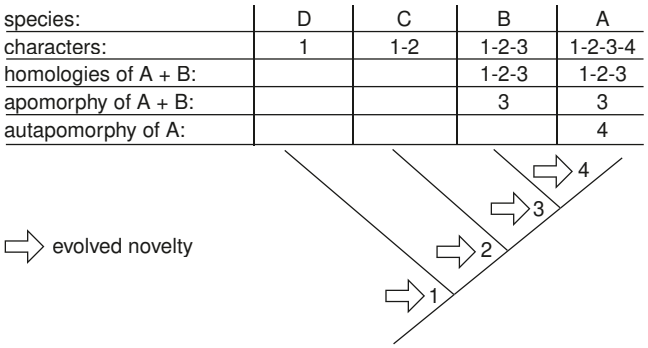


Fig. 76. The term apomorphy refers to a subset of a homology. Apomorphy and homology are not the same.

Attention: In practice, morphologists will often call the complete frame homology which contains novelties an "apomorphy", and the phylogenetically older state a "plesiomorphy". This allows a short, economical formulation. However, this usage is inaccurate and it is recommended to name the singular novelty (the detail) precisely, because a frame-homology always contains plesiomorphies as well as novelties: a "mammal lower jaw" may be called an "apomorphy" of mammals, however, the real novelties are not the teeth and bones as such but details of shape, position and ontogeny. – A morphological novelty can be the result of numerous mutations which occurred in a specific stem lineage. The delimitation of the stem lineage corresponding to a hypothesis of apomorphy follows from the corresponding sistergroup relationship (see Figs. 62, 64). – It is a mistake to call a species or a clade "plesiomorphic", because a species is not a character state and furthermore each organism is a mosaic of conserved and variable characters. The so-called "plesiomorphic species" are in reality species that retain a larger number of plesiomorphies and usually these organisms belong to a lineage that branched off early from a sistergroup with more derived characters. Do not use the expression "primitive" because of the negative connotation that is inappropriate. The species are "less derived" or "ancient".

The relationship between the terms "apomorphy" and "homology" has sometimes been misunderstood. Some authors think that the terms "apomorphy" and "homology" are synonyms (Patterson 1982, De Pinna 1991, Nelson 1994) and ignore that an apomorphy is a special homology, but not every homology is an apomorphy (Fig. 76).

An apomorphy occurring only in one species or in the basic pattern of a terminal taxon is called an **autapomorphy**. In Fig. 76 it can be seen that no statements on relationships are possible with the discovery of autapomorphies of a terminal species (character 4 of species A). Autapomorphies are **trivial characters**.

An apomorphy occurring in sister taxa (A and B in Fig. 76) which is obviously missing primarily in all other taxa (outgroup taxa), implying that it appeared in the stem species of the sister taxa A and B for the first time, is called a **synapomorphy** (Fig. 73; term of W. Hennig 1953, 1966). Only these characters can be evaluated as evidence for sistergroup relationships.

The state of a frame homology in the period before the evolution of a novelty is "**plesiomorphic**" *in relation to this apomorphy*. A plesiomorphy present in sister taxa is a **symplesiomorphy**. Symplesiomorphies can also occur in other organisms, including extinct or unknown species, and are not suitable as evidence for a sistergroup relationship.

Series of character states

The modifications of a frame homology occurring during the course of time can be lined up to a chain of chronologically successive character states. Such chains are **morphological series** or, for any type of character, **transformation series** (example: Fig. 77).

The assessment of the chronological sequence in which the novelties occurred historically is the

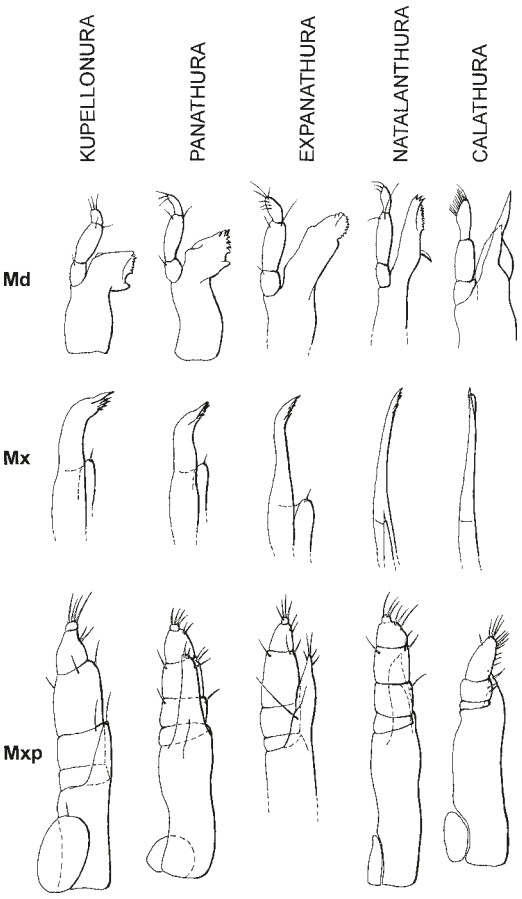


Fig. 77. Example for a transformation series (a "morphological row"). Evolution of mouthparts within the Anthuridea (marine isopod crustaceans). The Hyssuridae *(Kupellonura)* are carnivorous and have cutting mandibles (Md) and gripping maxillae (Mx), whereas at the end of the series the mouthparts of the specialized Paranthuridae (shown for a species of *Calathura*) can be seen, which are used to pierce through the cuticle of other arthropods to suck their body fluids. Mandibles (Md), maxillae (Mx) and maxillipeds (Mxp) respectively can be considered to be frame homologies, whose details were modified during the course of evolution.

determination of character state polarity (see ch. 5.3). This term is used when the changes observed within a frame homology (Fig. 77) are described. Often stepwise changes can be demonstrated without knowing the polarity of the series. In these cases it is not known which of the ends of the chain is the older state.

Homoplasy

It must not be forgotten that the terms "homology", "apomorphy", "plesiomorphy" always name hypotheses of which we hope that they correspond to real facts. In practice, hypotheses often turn out to be "incorrect", i.e. they cannot be verified and are contradicted by good evidence. This is especially noticeable when several hypotheses of apomorphy support incompatible groupings. Such incompatible characters are called "**homoplasies**": these are *a priori* potential hypotheses of homology which in a dendrogram are distributed as analogies or convergences (Fig. 78), but not as homologies. The hypotheses of homology appear to be incorrect. Note that a homoplasy is not necessarily always an analogy! A homoplasy can be:

- A real homology in an **incorrect phylogenetic** tree, in which the character occurs as apparent analogy.
- An **apparent reversal**, which is a real plesiomorphy on the wrong topology.
- A **real analogy**, convergence or parallelism that evolved through independent events and is mapped on a correct phylogenetic tree. Due to its lack of complex structure it cannot be distinguished from a homology and has been coded as homology.
- An analogy that originated from back mutations (**reversals**) and is recorded in the correct phylogenetic tree, and which cannot be distinguished from a homology.
- However, an **incorrect hypothesis of analogy**, which is based on an erroneous interpretation of a homology, will not have the distribution of a homoplasy, because each single character will be coded with a different number.

Primary and secondary homology

This distinction was introduced by De Pinna (1991). Primary homology is a hypothesis based on identity of details found in features of two or more organisms. Secondary homology is a hypothesis based on congruent character distribution in a tree topology (the cladistic homologization, see also ch. 5.3.3, 6.1, 6.1.10). Secondary homologies may be primary homology hypotheses confirmed after tree construction or new homologies that were not identified during data matrix compilation.

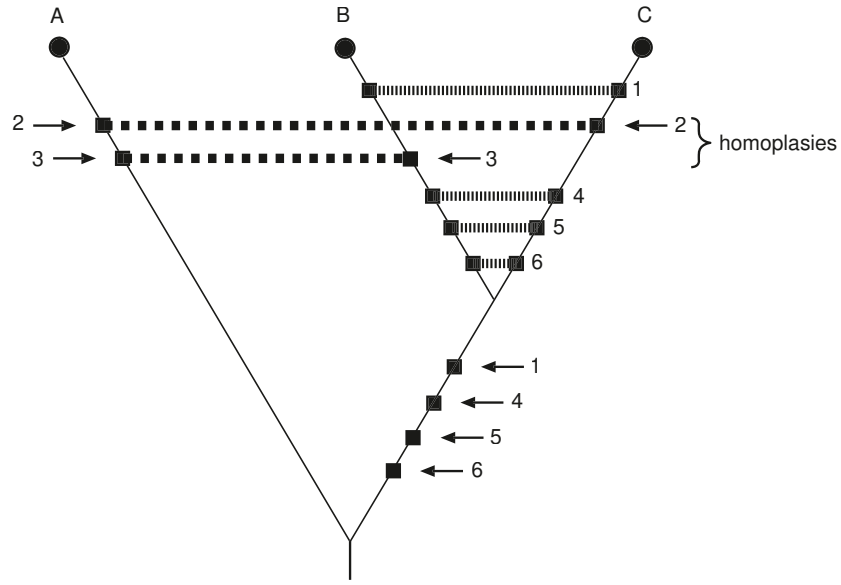


Fig. 78. Explanation of the term "homoplasy": most of the characters (1, 4, 5, 6) support the depicted topology, they are considered to be potential synapomorphies of the taxa B and C in the most parsimonious tree. However, the characters 2 and 3 are incompatible with this topology. As long as it is not known which of these contradicting characters are homologous with greater probability, i.e. when characters are unweighted and the topology may be incorrect, the incompatible characters (here 2 and 3) are neutrally called "homoplasies" (and not "analogies" or "convergences").

Convergence: non-homologous similarity which evolved due to adaptation to the same environmental conditions.

Analogy: non-homologous similarity which evolved by chance.

Homoiology: convergence which evolved from homologous organs.

Homology: genetically fixed information or the expression of such information, which has been inherited from a common ancestor of those organisms showing the character.

Parallelism: a term similar to homoiology, but referring to a parallel series of modifications. In most cases the evolutionary steps are not known and it makes no sense to distinguish between homoiology and parallelism.

Frame homology: a group of details forming a complex homologous pattern (character) or being physically combined. Within such a pattern not all of the details have to be homologous in different organisms.

Detail homology: small part of a complex character (a frame homology) that can be homologized in different species.

Apomorphy: evolutionary novelty which originated as result of mutations or gene transfer in populations of the stem lineage of a monophylum (= a new detail homology). Or (a second usage of the term): a frame-homology in which evolutionary novelties occur.

Autapomorphy or **trivial character:** apomorphy of a terminal taxon. Such characters are not informative for the reconstruction of phylogenetic relationships with parsimony methods.

Synapomorphy: homologous evolutionary novelty which can be used as evidence for a sister-group relationship and that evolved in the last common stem species of these sister taxa.

Apomorphic character state: a frame-homology composed of plesiomorphic and apomorphic detail homologies.

Plesiomorphy: homologous character (or state of a frame homology) in a state prior to the origination of an evolutionary novelty. A sistergroup relationship cannot be substantiated with it because this character state may also occur outside the considered monophylum or because the group bearing this character is para- or polyphyletic.

Symplesiomorphy: plesiomorphy occurring in sister taxa.

Homoplasy: term from the terminology of cladistics. It names a character whose distribution in a cladogram is not compatible with a hypothesis of homology. The term does not imply a decision in favor of a hypothesis of homology or analogy.

Characters state series or **transformation series:** chain of subsequent changes within a frame-homology. The chain can be reconstructed even if the polarity of character states is not known.

Polarity of a character series**:** chronological order for a series of evolutionary character state changes.

4.2.3 Forming groups with different classes of characters

Groups with convergent characters: These groups probably have no close phylogenetic relationships as long as no shared apomorphies are known. When the species evolved from ancestors belonging to different monophyla these groups are called **polyphyletic** (Fig. 50). When an Australian marsupial mole *(Notoryctidae)*, an animal living subterraneously, hunting insects and worms, reminds us of moles *(Talpidae)* of the northern hemisphere due to its cylindrical body, reduced eyes, short and strong digging legs, this is certainly an interesting observation for evolutionary biologists and ecologists: these animals belong to the same type of life form and have very similar ecological requirements. Such convergences cannot be used to substantiate phylogenetic relationships. The phenomenon of convergence should not be ignored, because these characters may be misleading and systematists have to distinguish between homologies and convergences.

Groups with plesiomorphic characters: When the distinction of groups of organisms is founded on the presence of plesiomorphies, the resulting taxa are **paraphyletic** or possibly polyphyletic. Reptiles, for example, are those amniotes (= Tetrapoda without amphibian life-cycle) which have neither feathers nor mammalian hairs. The shared characters of the "Reptilia", namely eggs with shells, the horny, usually scaly skin which is poor in glands, the presence of neck ribs and other characters are plesiomorphies. The exclusion of mammals and birds from the "Reptilia" has the effect that this group is paraphyletic; inclusion of all taxa would make it the same as the Amniota.

Groups with apomorphic characters: We can only talk about apomorphic characters when a homology is an evolutionary novelty. When the delimitation of a group is supported with this type of character, implying the hypothesis that there was a last common stem species that had this character for the first time, this group is a **monophylum** if all descendants of the last common ancestor is included in the group. However, not only carriers of this character belong to the monophylum, but possibly also other descendants of the stem species where the apomorphy has been reduced or modified secondarily.

Remember that it is necessary to distinguish between the **being** and the **identification** of a monophylum. A monophylum is in the first place an intellectual concept that implies a hypothesis on the existence of a common ancestor (ch. 2.6). The group is monophyletic because we assume that its members have a common descent, not because it shows certain characters. Descent is not a character, but a historical process. Apomorphies are identified evidences for the existence of this process. Therefore, systematists search for characters supporting a hypothesis of monophyly. Snakes, blindworms (Anguinae) or whales are classified as Tetrapoda although they do not have four walking legs. The existence of organisms that can be united in a monophylum Tetrapoda is the result of descent from a last common ancestor, the observation of their walking legs and other characters are the motive for the distinction of this monophylum.

Having refined the definition of "apomorphy" we can now discern more consciously between "***homology signal***" and "***phylogenetic signal***": homologies are of course traces left by evolutionary processes and they contain phylogenetic information. However, to reconstruct phylogeny we need homologous *apomorphies*. It is therefore important to distinguish noise and signal in a dataset, and to find out if signal is composed of apomorphic states. Signal visualized with spectra, for example, is composed of homologies (Fig. 154). It depends on taxon sampling whether all of the signals are composed of apomorphies or if plesiomorphies support paraphyletic groups (see also ch. 6.3.3)

Homology signals: non-random patterns of similarities found in the morphology or in molecules of organisms that can be explained with common descent (in constrast to noise). Homology signals can be composed of apomorphies *and* plesiomorphies. Signals composed of plesiomorphies are misleading.

Phylogenetic signals: these are homology signals composed only of apomorphies. They can be used to reconstruct phylogeny.

4.2.4 Homologous genes

Further terms have been coined for the homology of genes (Patterson 1988):

Paralogy: Homology of duplicated sequences occurring in one organism (e.g., α- and β-hemoglobin, HOX-genes, phytochromes, hemocyanines, RNA-polymerases, etc.). Paralogy corresponds to the **homonomy** of duplicated morphological structures. It is methodologically important to distinguish paralogous genes. When phylogeny is calculated on the basis of the comparison of paralogous genes, the divergence seen in a gene tree may not correspond to the divergence of species when a gene duplication occurred long before the speciation event (see Fig. 7). Paralogous genes may differ significantly in structure and function and can evolve with different substitution rates. Genes coding for repetitive organs must not necessarily be duplicated and paralogous, they are probably in many cases exactly the same single genes that are activated in different parts of a body.

Orthology: Homology of sequences which did not originate in gene duplication but due to speciations. The true gene tree has the same topology as the species tree.

Xenology: Homology of sequences of unrelated organisms which originated from a horizontal gene transfer (see ch. 2.1.1).

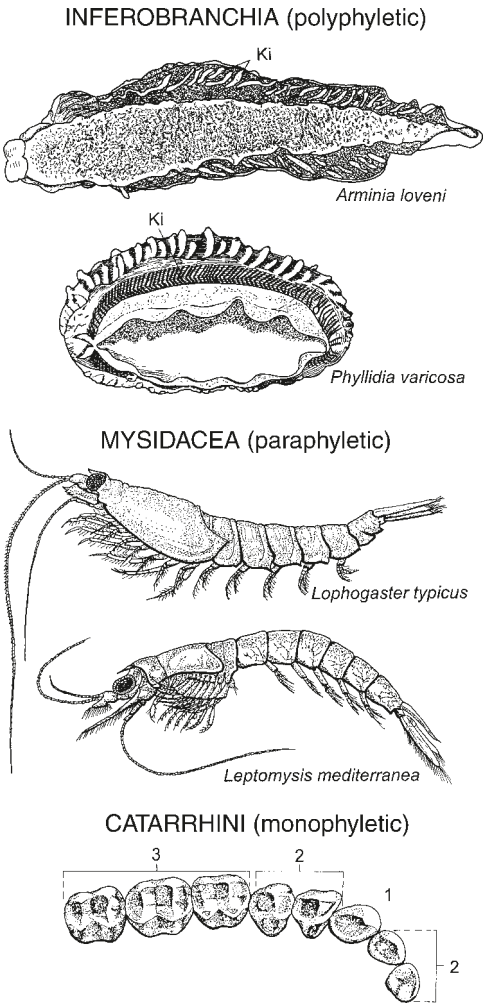


Fig. 79. Examples for the distinction of groups using non-homologous characters, plesiomorphies, or apomorphies. The "Inferobranchia" (Gastropoda: Nudibranchia) have secondary gills (*Ki*), which develop from folds of the mantle epithelium. These gills are a convergence of the Arminidae and Phyllidiidae, the taxon Inferobranchia is polyphyletic. – The "Mysidacea" are primitive Peracarida which have a plesiomorphic shrimp-like habitus. The Mysida share apomorphic characters with other peracaridan crustaceans, the taxon Mysidacea is therefore probably paraphyletic and founded on symplesiomorphies. – The monophyly of the Catarrhini (old world monkeys, including humans) can be substantiated for example with the tooth formula (1 premolar is missing) and the form of the molars (in the ground pattern tendencies to bilophodonty) (depicted is the dentition of a fossil species of *Dolichopithecus*; after Szalay & Delson 1979). 1, 2, 3,: number of tooth types (from left to right: posterior molars, anterior molars (premolars), canine, incisors).

4.3 Principles of character analysis

In this chapter first principles serving the identification of homologies will be distinguished. In later chapters methods needed for this purpose will be introduced in more detail. We first of all are interested in the type of character analysis which should be performed *prior* to the reconstruction of phylogenetic trees.

In ch. 1.3.7 it has been explained that characters are mental constructs which are based on perceived similarities. The further analysis of characters presupposes the real existence of corresponding details and that our perception is not defective.

The decisive step of character analysis is the distinction between homologies and chance similarities (convergences, analogies). According to the comments in ch. 1.4.5 a statement of probability for the alternative "homology or analogy" can concern either the amount (complexity) of information present, which can be identified as trace of historical processes by a suitable receiver (probability of cognition), or the statement refers to the probability of natural processes which produced the characters (probability of events).

Therefore two very different approaches can be used:

a) **Phenomenological character analysis**: estimation of the probability that the identical details of two characters stem from a common source (**analysis of patterns, estimation of the probability of cognition**) (Fig. 80, see also Fig. 87 and Fig. 139).
b) **Process-dependent (modelling) analysis** of the probability that a character (character state) originates or that a character is transformed into a new state (**reconstruction of processes, estimation of the probability of events**) (Fig. 81).

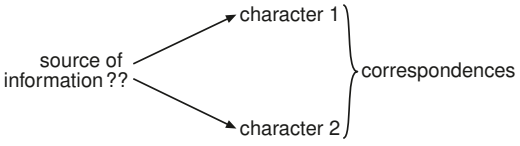


Fig. 80. The comparison of patterns is used to estimate the probability of the presence of a homology.

This differentiation reminds of R. Riedl's distinction between the "**act of explaining**" and the "**act of cognition**" (Riedl 1975). When a character is recognized as the result of a reconstructed historical process, one explains how it evolved. It is something totally different to ask whether the observed object really represents a trace of historical events or not.

The phenomenological character analysis is useful to estimate the *probability of homology* for observed similarities. It will be introduced in ch. 5.

The process-dependent analyses serve the estimation of the *probability of events* for the transformation of characters (ch. 7). For this purpose, parameters which influence the evolution of characters have to be considered (see ch. 2.7.1). This is rarely attempted for morphological characters (compare ch. 2.7.1). For the evolution of sequences these variables are the reconstructed ancestral sequence and parameters that influence the number of changes (substitutions, insertions, deletions, translocations, etc.) that occurred per unit of time. Usually only the rate of nucleotide substitutions is estimated. Large insertions and other unique mutations are hardly predictable. The different types of substitutions can be estimated or modelled but the process cannot be observed directly. It has to be taken into account that one axiomatic assumption is the basis for all model-dependent methods: the substitution processes have to be stochastic. More on this subject in the chapters 7, 8, and 14.1.

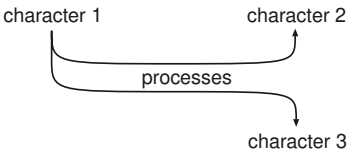


Fig. 81. When statements on the most probable course of a process are made, assumptions on a common starting (ancestral) state of two terminal characters can be deduced, even when these terminal characters do not share identities.

4.3.1 Processes and patterns, or what we can learn from Leonardo's Mona Lisa

Discussions with students proved that it is not easy to understand which implications the consideration of processes and patterns has for character analyses, and which are the patterns we are talking of. A quick look at the analogy of the duplicated Mona Lisa helps to understand what we are doing when we consider patterns and processes. This famous painting by Leonardo da Vinci has been copied often by other painters and cartoonists. Those who are familiar with the original will be able to recognize the smile of Mona Lisa even in a very simplified or disproportionate cartoon. In Fig. 82 one can see the original and three copies, which have been produced from the original by distortions and change of colours. In the last copy (Fig. 82D), the original can hardly be recognized. This example corresponds to the problem we have when we try to identify homologies and it has to be interpreted as follows:

- The original by Leonardo corresponds to an ancestor or an ancestral organ.
- Each copy corresponds to a descendant or to an organ of a descendant.
- Comparing original and copy, the original as well as each copy correspond to a complete *frame-homology*.
- Each detail corresponds to a *detail homology* whenever the same detail can be seen in at least two of the pictures.
- The process of modification of the original corresponds to the evolutionary process.

For a character analysis the following questions have to be settled:

- Is a picture really a copy of the original? Or, translated to biology: is an organ of a species really homologous to an organ of the ancestor?
- Are two pictures really copies of the same original? Or: are two organs of different species homologous?

There are two possibilities to answer these questions.

a) **Analysis of the process** (reconstruction of the course of the copying process): to use assumptions about the copying process it is necessary to take into account parameters of this process. The

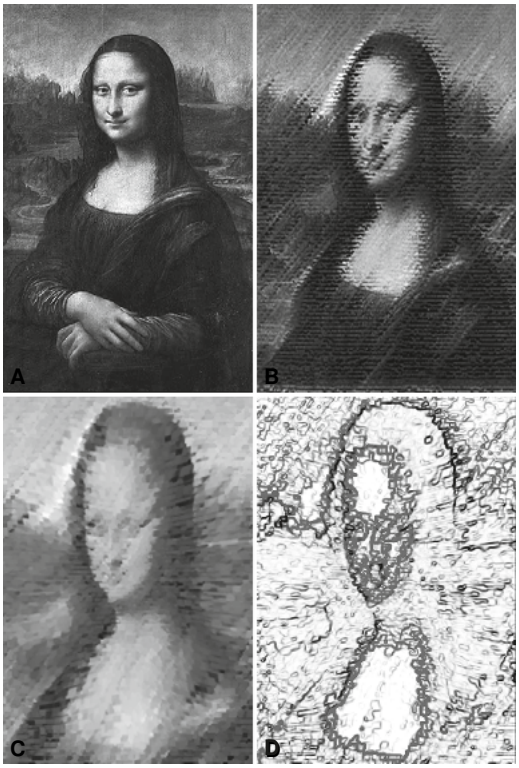


Fig. 82. An original (picture A) and its modified copies (B-C). The identification of homologies corresponds to the finding that two pictures are copies of an original. When the *process* of copying cannot be reconstructed, the identification of copies depends on the presence of congruent details. When a picture (as in Fig. 82D) has really been produced as copy of an original, the original is only identified when enough details of it were retained in the copy. When two copies (e.g., C and D) are compared, the statement that there must exist a common original corresponds to a homology statement.

original cannot be reconstructed merely by stating that "there was a copying process". Neither is it helpful to simply state that evolution occurs: these assumptions on the existence of a process do not allow a reconstruction of a ground pattern or of an ancestral character. Whenever the process has been completely documented, a copy can be transformed pixel by pixel backwards to the original state (this may be done with the computer using the command "undo"). This means for the case of a modified picture that it should be known which shifts of points occurred and which pixels have been deleted or added. Such a reconstruction is the proof that a copy is "homologous"

to the original. It does not have to be considered in this case which other copies exist additionally. Or, translated to biology: *the phylogenetic tree does not have to be known*, the path between copy and original will be reconstructed with process parameters. The analysis of the process also allows a statement on homology even if we cannot recognize by eye in the copy (Fig. 82C) features of the original (Fig. 82A).

It can be envisioned that lawful relationships which allow the reconstruction of *a part of the copying process* may be discovered by comparison of several conjectural copies. In this way a statement on homology becomes possible even when a *recording* of the process does not exist. The copies Fig. 82C and Fig. 82D show for example common features compared to Fig. 82B (elongation of the head, shortening of the body), which allow to identify C and D as result of the same change (or, if the direction is not known, of the same process separating {C,D} from B). When the distortion seen in C and D is undone, proportions of the head can be transformed to those of picture B. The statement "picture B and picture D have a common original shared with picture C" corresponds to a statement of homology and depends on the assumption that a process of distortion of head proportions occurred. Such a statement is also possible when the original (Fig. 82A) is not known. It is important to note that such a statement of homology is based on a *backwards reconstruction of the process*. Therefore, the probability that two copies are homologized correctly depends on whether the assumptions on the course of the process derived from existing copies correspond to the real series of events. We have to evaluate the probability that a specific process such as the transformation from picture A to picture B or C or D really occurred. Analyses of processes will be discussed in chapters 7 and 8. They play an important role in molecular systematics. Whenever parameters that influence the selection of morphological characters are well known (as in the case of the bills of Darwin finches, compare Fig. 36 and text referring to it), evolutionary processes can also be modelled for phenotypes. In the practice of systematics, however, this opportunity is hardly ever given.

b) **Pattern analysis**: If the copying process is not documented or not reconstructable, homologization is nevertheless possible by comparison of details. The more details are congruent in two pictures, the greater is the probability that we can recognize features of the original in a copy. In Fig. 82B and 82C the distribution of light and dark areas is very similar, the contours of the background at eye level is similar, the turning of the shoulders is the same, etc. With these observations the assumption can be justified that both pictures have a common original or that C is a copy of B (or vice versa). This way of substantiation does not require any assumptions on the course of the copying process. A "transformation series" can be postulated, arranging the pictures in such a way that those with greater similarity are neighbours: A↔B↔C↔D. It cannot be determined without additional information in which direction this series has to be read. But the series enables to homologize D with A, even though possibly nobody would be able to recognize the Mona Lisa with certainty seeing only picture D. During pattern analyses the probability that details of two pictures can be recognized correctly as being identical is evaluated.

Note that the probability of identity of details increases the probability of homology for the whole picture (the frame homology; see ch. 5.1). And vice versa: if it is true that a picture is a copy of an original, the probability of homology is also high for identical details. The differentiation between frame and detail homologies is necessary for methodological reasons, because without it a homologization using a phenomenological pattern analysis is not possible.

4.4 Delimitation and identification of monophyla

A monophyletic group can only be recognized because the historical existence of common ancestors is a fact. All individuals derived from a certain ancestor carry specific genes and mutations which were already present in this ancestor. The opportunity to identify a monophylum results from this fact. Because immediately after a speciation event the diverging daughter populations are at first very similar, the identification of an isolated daughter population is difficult at this point of time. The *being* (fact of existence), however, is independent from the chance of *recognition*. We have to distinguish:

– The cause for the existence of a certain monophylum (a specific speciation event).
– The motive for recognizing a certain monophylum (presence of an evolutionary novelty).

4.4.1 The delimitation

There exist more monophyletic groups in the four dimensions of space and time of a phylogenetic tree than there were speciations in the course of earth's history (see ch. 2.6, Figs. 28, 83). Only very few of these monophyla can still be recognized today, because most of the organisms that once lived on earth are not preserved as fossils. An even smaller portion gets proper names for the purpose of scientific communication. In order to avoid misunderstandings it has to be pointed out to which monophylum a proper name refers.

Theoretically there are four possibilities to delimit monophyla that comprise different groups of organisms. In each case a specific time level in which the boundary to other monophyla lies has to be defined.

1) A monophylum can be distinguished from other ones with reference to the **last common stem species**. In practice, the indirect identification of the unknown stem species is usually achieved by uniting known species in a group whose last common ancestor is inevitably the stem species (case 1 in Fig. 84). This determination excludes older stem lineage representatives from the monophylum.

2) Select one or more unique derived characters (**apomorphies**) which occur only in members of the monophylum (case 2 in Fig. 84). The phylogenetically first member of this group is the individual that possessed this character or set of characters for the first time in history. Stem lineage representatives which do not

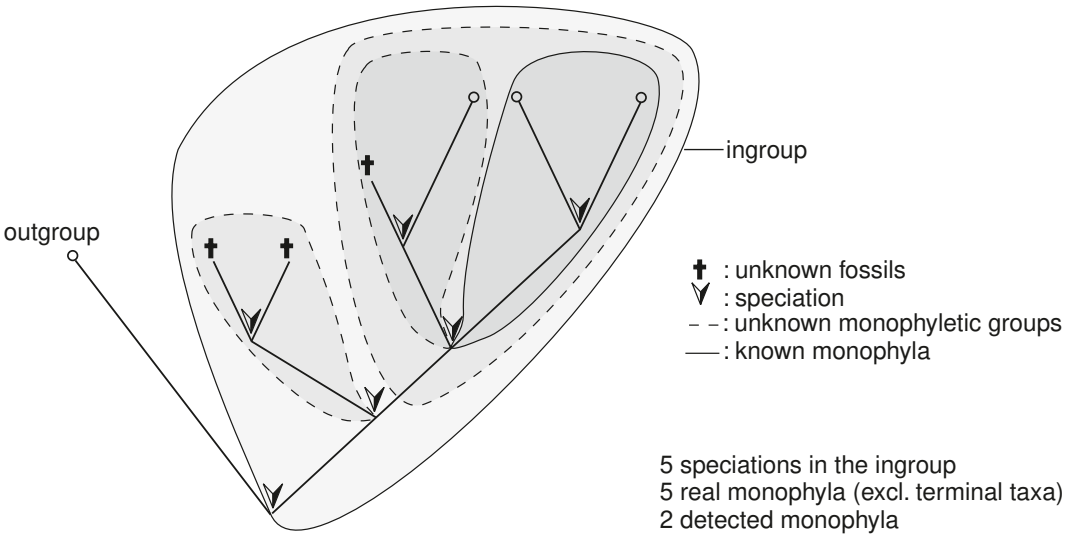


Fig. 83. Existing and discernible monophyla.

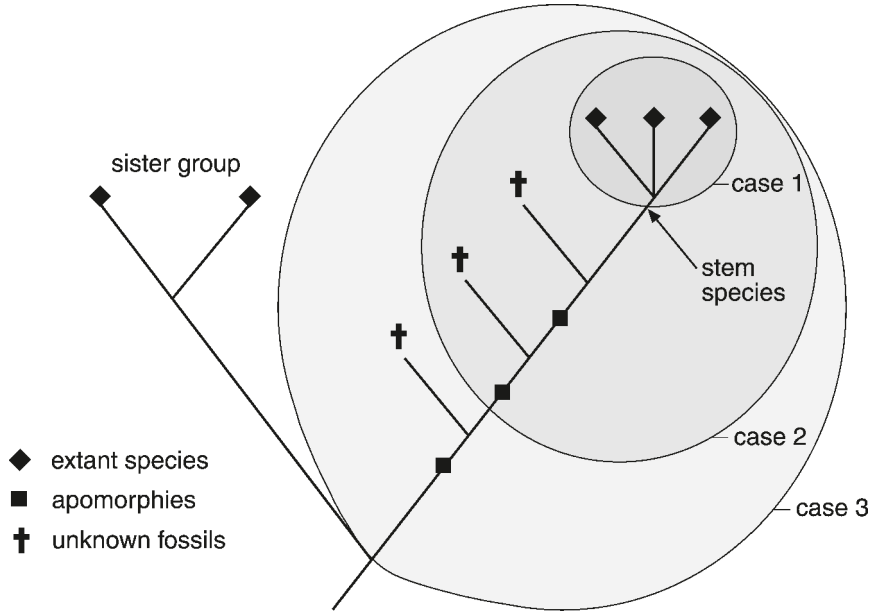


Fig. 84. Different ways of delimitation of monophyla. Monophyla can be delimited naming a last common ancestor, one or more apomorphies, or a sistergroup.

yet possess these apomorphies have to be excluded from this taxon. In Fig. 84, however, it is visible that definitions of monophyla are less dependent of our knowledge about the fossil record and unequivocal when they refer to the topology of a phylogenetic tree and not to characters, because fossils may be unknown or they may not preserve the relevant apomorphies. And, apomorphies are less suited for the delimitation of monophyla, because they originate stepwise and in a mosaic-like fashion, possibly even within the time of existence of a single species. When the characters "feathers" and "pygostyl" are defined to be constitutive for the taxon "Aves", the genus *Archaeopteryx* is excluded, because it lacks a pygostyl. If, however, the tarsometatarsus is chosen, *Archaeopteryx* is a member of the taxon "Aves".

3) **Name the sister taxon** (case 3 in Fig. 84). This demarcation has the advantage that even with increasing state of knowledge on, for example, characters of fossils and with changing views on the homologization of single characters the point of reference, namely the last common ancestor of sister taxa, is retained in the dendrogram. Stem lineage representatives are not excluded (compare the term "panmonophylum", ch. 3.5, 3.6).

4) **Select two terminal taxa**. The monophylum is defined with the last common ancestor of these taxa. A disadvantage of this method is that inadvertedly some terminal taxa may be excluded which are similar to the other ones but derived from an older ancestral species.

In practice, the **first stem species of a monophylum** that includes the stem lineage of a crown group is usually not known. But theory allows an unequivocal identification of the place of a stem species in a tree. Since *according to its definition* a phylogenetic species stops to exist "when it splits into daughter species" (see ch. 2.3), the stem species from which two sister monophyla evolved cannot be included in one of these monophyla. Otherwise the same stem species would belong to two monophyla. The phylogenetically oldest species which can belong to a monophylum is the one whose (conceptual) existence starts immediately after the splitting of the sister taxon.

4.4.2 The identification

Even though there are different possibilities to delimit monophyla from other groups in a phylogenetic tree, there must be a motivation to group species. In phylogenetic systematics the distinction of groups can be substantiated with apomorphies that could be identified either directly or indirectly via distance or other quantitative measures. These apomorphies should be evolutionary novelties of high probability of homology, which are assumed to have been evolved in the stem lineage of the monophylum (ch. 4.2.3). Methodological principles needed for the identification of single homologies and the distinction between plesiomorphies and apomorphies are introduced in ch. 5. Using indirect methods that rely on probabilities of character transformations, it is necessary to estimate the probability that a number of novelties evolved. Using likelihood methods this probability is estimated for the whole set of characters in one single analysis using assumptions about substitution rates (see ch. 8).

4.4.3 Recommended procedure for practical analyses

1) The monophyly of a group of known (recent and fossil) species is always substantiated with apomorphies. Using discrete putative apomorphies, these should have a high probability of homology and one should be able to name them as single characters. Using transformation probabilities, the assumptions about substitution rates should be based on excellent empirical data, the probability that frame characters are homologies (e.g., orthologous genes, alignment positions) should be high, and the probability that similarities are plesiomorphies (see ch. 6.3.3) or convergences (e.g., due to parallel shifts in base composition) should be low.
2) The sister group should be identified.
3) Both adelphotaxa inevitably include the corresponding stem lineage representatives, independently of whether the latter already show all apomorphies of the known species or not. Therefore the adelphotaxa each are in relation to their corresponding crown groups of extant species *panmonophyla* (ch. 3.6).
4) Cases in which it is tradition to use a name for the crown group alone, either a new name has to be given to the panmonophylum, or the more comprehensive taxon has to be called *pan*taxon (Panmandibulata: Fig. 65).

Nature is often complicated and difficult or impossible to fit into our systems of terms: in rare cases there exists horizontal gene transfer between monophyla, especially through processes which occur with the evolution of endosymbionts (see also chapters 2.1.1, 2.1.4). Where hybridizations are possible, introgressions of genes may occur. In such cases it can happen that different evolutionary novelties which did not evolve in the same stem lineage can occur in a single organism.

4.5 Analysis of fossils

4.5.1 Character analysis

Fossils are traces of extinct organisms whose characters do not differ in any way in their ontological status from those of recent organisms. Problems occurring in practice, for example because fossils are badly conserved, are not a peculiarity of fossils: many recent species have been described so poorly that the state of knowledge is not better than for badly preserved fossils. For this reason special laws for the inclusion of fossils in a phylogenetic analysis are not required. It is, however, essential that characters coded in data matrices are homologies. Fossils provide very interesting data. They often allow

- to determine a **minimum age for characters and taxa**, or
- to close **gaps in transformation series** of characters. This is often the only way to homologize structures (frame homologies) that are very different in extant species.
- They allow to determine the chronological sequence of transformations ("what has been there first?").

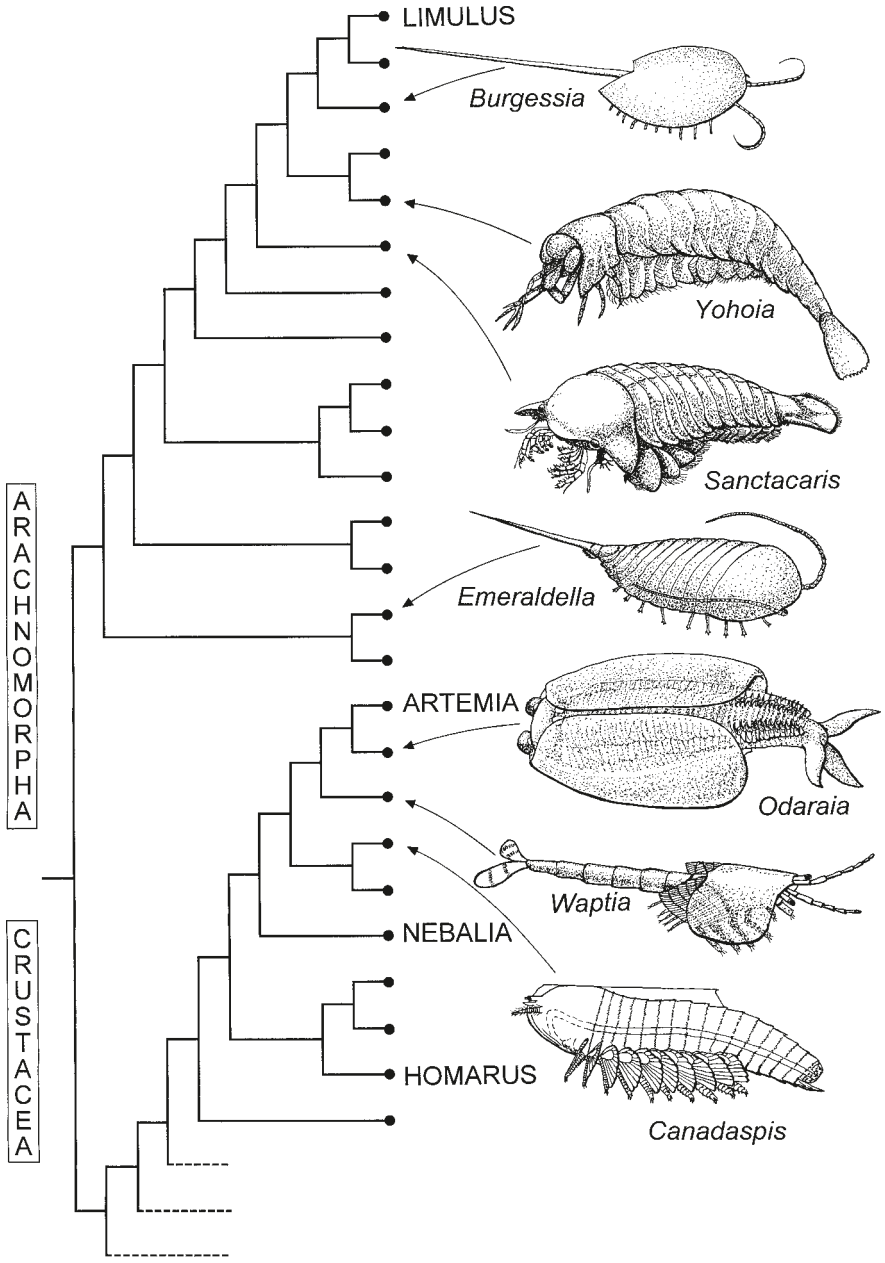


Fig. 85. Hardly justified assignment of fossils to taxa which were originally erected for recent species (topology according to cladistic analysis of Briggs et al (1992), illustrations after Briggs (1992) and Briggs et al. (1994)). According to Briggs et al. (1994) *Sanctacaris* belongs to the Chelicerata. However, appendages similar to chelicera or other specific apomorphies are not present. Additionally this fossil also has antennae, which are lacking in Chelicerata. The segmentation of the body (prosoma with 6 walking legs, opisthosoma with 11 segments and leaf-like exopods) support the idea that it may be a stem lineage representative of the Chelicerata. – Neither the body segmentation of *Burgessia* nor any other known character resemble features of the Chelicerata. – *Odaraia*, *Waptia* and *Canadaspis* are assigned to the Crustacea, although neither the presence of apomorphies of the Mandibulata (such as the specific differentiation of mouth parts and of the second antenna) nor possible apomorphies of Crustacea have been demonstrated. This systematization is unfounded.

The information relevant for phylogenetic analyses should be gained through careful character analyses (ch. 5) in the same way as with recent organisms.

Fossils can be very valuable for the evaluation of characters. The fact that similarities of recent species are based on convergence can be proven with fossils (Willmann 1990). For example, the females of recent Dermaptera as well as those of {Mantodea + Isoptera + Blattaria} have a reduced ovipositor resting in a genital chamber. Hennig (1986) considered this character a possible synapomorphy. In fossil Dermaptera, however, a long ovipositor is present. The reduction in recent species is a convergence to {Mantodea + Isoptera + Blattaria}. Fossils can show synapomorphies which due to later modifications are invisible in recent species. For example, the shape of bills of fossil flamingos is the same as in plovers. This similarity is not seen in recent species.

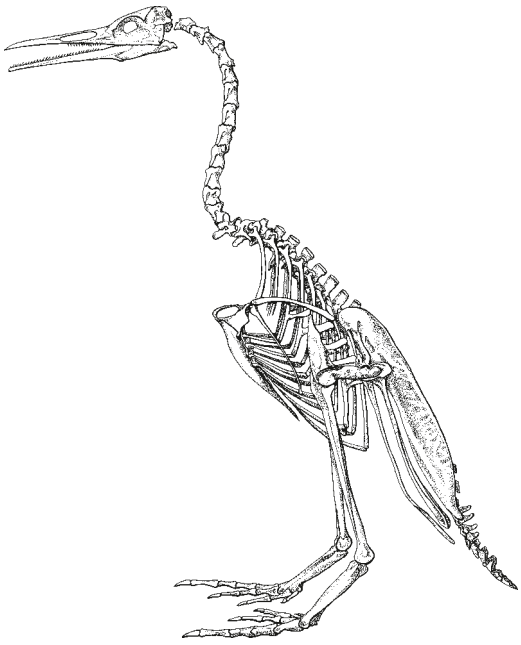


Fig. 86. Autapomorphies of a stem lineage representative: *Hesperornis regalis* is a primitive bird from the Cretaceous which still possessed teeth. Autapomorphies of this group of flightless marine birds are the reduction of wings, reduction of the system of air sacs, increased pelvic length, and the patella forms a process adapted as attachment site for large muscles.

4.5.2 Transformation series of populations as evidence for monophyly

When series of consecutive fossil populations with known chronology are present, the fossils are documents for monophyly even when no apomorphy is found. Such a series can be a documentation of the sequence of speciation events (Willmann 1985, 1990). However, a complete series as the one in Fig. 22 is rarely preserved. The additional information concerning the time level in which a fossil occurs becomes important

- to calibrate molecular clocks (ch. 2.7.2.3),
- or when aspects of natural history are considered which are beyond phylogenetics in the stricter sense, but important for the check of the plausibility of a hypothesis. Examples are the estimation of the time available for the evolution of diversity or the correlation with geological events (orogenesis, origin of lakes and islands, etc.)

Singular fossils which (in inaccurate phrasing) are considered to be "predecessors" of a recent monophylum must have evolutionary novelties which only occur in the considered monophylum. In the past some paleontologists have succumbed to their "intuition" and placed fossils in superficially similar recent taxa (example: Fig. 85). Only an apomorphy of high probability of homology can substantiate such an assignment. Such "predecessors" of recent monophyla should be called "**stem lineage representatives**" (see ch. 3.5):

It is highly improbable that of all the many species belonging to a stem-lineage and branching from it a fossil belonging to a directly ancestral population of recent organisms was preserved. Therefore the term "stem lineage representative" has to be preferred over the term "ancestral species". The fact that a "stem lineage representative" belongs to a side branch of the stem lineage can often be shown with autapomorphies of the species (Fig. 86). Fossils can, but do not necessarily have to belong to a terminal species.

5. Phenomenological character analysis

Each phenomenological character analysis has to start with the analysis of the properties of an organism. The more carefully an organism is examined, the more of its properties can be detected. Frequently observed properties of physical objects are, for example, the shape of a structure and of its components (also the histological structure of an organ, or the shape of organelles within a cell), chemical composition, positional relations, specific physical or chemical interactions with environmental factors (e.g., refraction of light, quality of reflected wave lengths, colour). Properties of macromolecules are chemical composition, position (sequence) and type of bond between monomers; behavioural properties are, for example, the specific sequence of movements and the circumstances that trigger a specific reaction. That identity of two patterns has been found is often expressed with the statement "two organisms have the same character". The main objective of phenomenological character analysis is the detection of homologous identities, or, to be more accurate, the discovery of empirical evidence that supports a hypothesis of homology.

The phenomenological method does not go beyond the observed phenomena (ch. 1.4.6). In the following, those methods are called phenomenological which start with the observation and comparison of properties of individual organisms and evaluate these observations without reference to assumptions about those historical processes that may have produced the visible identities and differences. In short, patterns are analysed when processes are unknown or not taken into consideration. These first empirical observations lead to a classification of characters and to the formulation of hypotheses of homology which initiate the hypothetico-deductive method in phylogenetic systematics.

5.1 The estimation of the probability of homology and character weighting

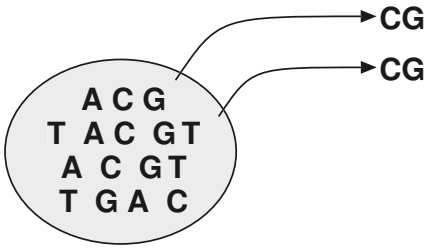
We have to distinguish the following parameters:

- the *historical cause* for the occurrence of homologies (namely descent and transmission of genome copies),
- the *effect* caused by descent and inheritance (similarity seen in organisms),
- the *motivation* to postulate homology (conspicuous similarity of structures which cannot be explained with convergence or accidental correspondence).

In the following sections the assessment of probability of homology (which is the motivation for postulating a hypothesis of homology) and the utility of this estimation in phylogenetic analysis (for weighting of characters) are introduced.

5.1.1 The probability of homology and criteria for its evaluation

Within the framework of phenomenological analyses the identification of a homology is a prerequisite for any proposal of a hypothesis of relationships. A statement of homology itself is again a hypothesis (compare ch. 1.3.7 and 4.2) which has to be substantiated. The hypothesis should not be considered to be a "fact", because this misinterpretation can lead to erroneous statements of phylogeny in cases when a statement of homology is based on an error. Many implausible results of cladistic analyses (see Fig. 54) are based on the erroneous belief that homologies listed in character tables are facts and do not require critical scrutiny. This misunderstanding concerns morphological as well as molecular characters. We have to estimate how **informative** characters are, or in other words, how probable it is that we can recognize correctly a homology-



$P = (1/M)^n$

P: probability that a word is constructed by chance alone
M: size of the alphabet
n: number of letters in the word
(assumption: all letters are equally frequent in the pool)

alphabet:	**4** letters (A G C T)	**36** letters
length of the word:		
2 letters:	**P = 0.0625**	**P = 0.00077**
10 letters:	**P = 1 x 10^{-6}**	**P = 2.7 x 10^{-16}**

Fig. 87. Estimation of probability of homology using the criterion of complexity relies on a probabilistic law.

relationship between two similar patterns. Phenomenological analysis does not strive to model the evolutionary process that shaped characters.

In ch. 1.3.6 it has been explained how "information" is quantified in informatics. Unfortunately, Shannon's formula cannot be used for our purpose in practice: in order to compare information content of different characters (patterns) with a universal measure ("bits"), we have to know the average probability for the evolution of a single detail, for example a single mutation, which has to appear in an individual and afterwards must be conserved in all following populations. This corresponds in the analogy of Fig. 87 to the probability that a certain letter is selected at random from a pool containing an alphabet of available letters (Fig. 87, explained below). As this probability cannot be estimated for morphological characters without detailed knowledge of all parameters relevant for the evolution of the organisms under consideration, it is futile to attempt a calculation of information content in "bits". For sequence data with estimated molecular clocks (calibrated substitution rates) this approach has not been proposed until now.

Furthermore, Shannon's formula requires that the "transmission" (thus, in the case of homology, the inheritance of homologous genes) proceeds free of interference, assuming that noise that may destroy information or create misleading patterns does not exist. From the point of view of a systematist, disturbance in the flow of information within reproductive communities results from genetic drift, due to which gene variants are lost, and from mutations which create analogies or substitute apomorphies. However, the fundamental law which has to be considered in homology research is the same as in the analogy of patterns composed of letters (Fig. 87): the more complex a character, and the more alternative types of letters are available, the more informative is the character, or, in other words, the lower is the probability of chance similarity between two similar patterns. For this reason it is convenient to know Shannon's notion of quantification of information.

The relationship between complexity and probability of chance similarity can be illustrated with the analogy of a machine selecting letters (Fig. 87): letters are selected at random from a pool to construct two words of the same length. It is assumed that the pool is inexhaustible and that letters occur in it with equal frequency. Then the probability **P** that two identical words are constructed by chance only depends on the length **n** of the word and the size of the alphabet **M**. The table in Fig. 87 shows which dramatic influence **M** and **n** have on **P**.

This analogy requires assumptions on the apparatus which selects letters: the selecting machine should not have a preference for specific letters, it works like a "fair die", and the frequency of letters in the pool is equal. However, should the machine be unfair and favour specific letters, the probability that identical strings are constructed by chance increases and is maximal when only a single letter is selected. The more complex the string, the smaller is the importance of fairness. In long strings and with a large alphabet, small deviations from the equal distribution of letters or a small bias in the selecting process can be neglected and it can be assumed that it is little probable that identical strings are produced by chance. A more general formulation for the probability that two words are equal by chance is

$$P = (\sum_{i=1}^{M} p_i^2)^n$$

In this formula p_i is the frequency of the individual letters. The formula in Fig. 87 is only a special case where p_i is assumed to be the same for all letters.

Note that the time factor or the *rate of the process* is not relevant in this context. A precise estimation of the probability of chance similarity requires knowledge on the structure of patterns, on the size of the alphabet, on the frequency of single letters in the pool and on the preferences of the selecting apparatus.

It is interesting to compare the probabilities for a string of **coding DNA** and for the corresponding **amino acid sequence**. For example, using the model with equal frequencies of letters and having 120 nucleotides, the probability of getting twice the same string by chance is $5.66 \cdot 10^{-73}$, while for a string of 40 amino acids and 20 different symbols the probability is $9.09 \cdot 10^{-53}$. This simple calculation shows that DNA sequences should be much more reliable for phylogeny inference.

While the comparison of sequences allows the estimation of frequencies p_i, morphologists usually do not have this possibility. Nevertheless, it can be assumed also for morphological characters that complexity is an indicator for probability of homology. On condition that the frequencies of different p_i-values are randomly distributed, it can be assumed that the most probable case is an increase of probability of homology with increasing complexity.

Evolution is the process selecting those elements of which characters are composed of, the "pool of letters" contains the number of alternatives that exist for the construction of molecules, organelles, cell types, tissue types, arrangements of organs. The alphabet of nature is very large, but in practice we use simplified operational alphabets to describe visible aspects of real organisms. The composition of the selected operational "alphabet" depends on the methods used to study patterns (e.g., biochemistry, cytology, anatomy). The information content varies with the operational alphabet. This phenomenon is well known to molecular systematists who use for the same sequences either a translated amino acid alphabet, a RY-alphabet or a AGCT-alphabet. We estimate the probability that patterns composed either of amino acids, or of purine- and pyrimidine-nucleotides, or of A, G, C and T evolved.

To understand the basis of homology research, we have to assume that during evolution a "possibility to choose between alternatives" exists. On principle, an organ needed for specific functions can be constructed in various ways with different modules of different origin. This assumption is testable: a retina, for example, can grow during ontogeny from epidermal anlagen (in many invertebrates) or from embryonic nerve cells (in vertebrates), it can be constructed as inverse or everse sense organ, photoreceptor cells record a stimulus with cilia or with microvilli, etc. Of course, the supply of modules is limited and depends on the raw material at hand. In eukaryotes, for example, the number of different types of cell organelles and of pathways to construct organic molecules limits the number of available modules. The number of available alternative modules is analogous to the size of the alphabet. As we generally do not know neither the number of possible alternatives for morphological characters nor the probability that a complex pattern evolved, an absolute value for the probability **P** that a character evolved cannot be estimated. But we can estimate **relative probabilities** considering these observations: the more complex a character is and the more alternative modules are known from nature, the higher is the probability that two identical patterns did not

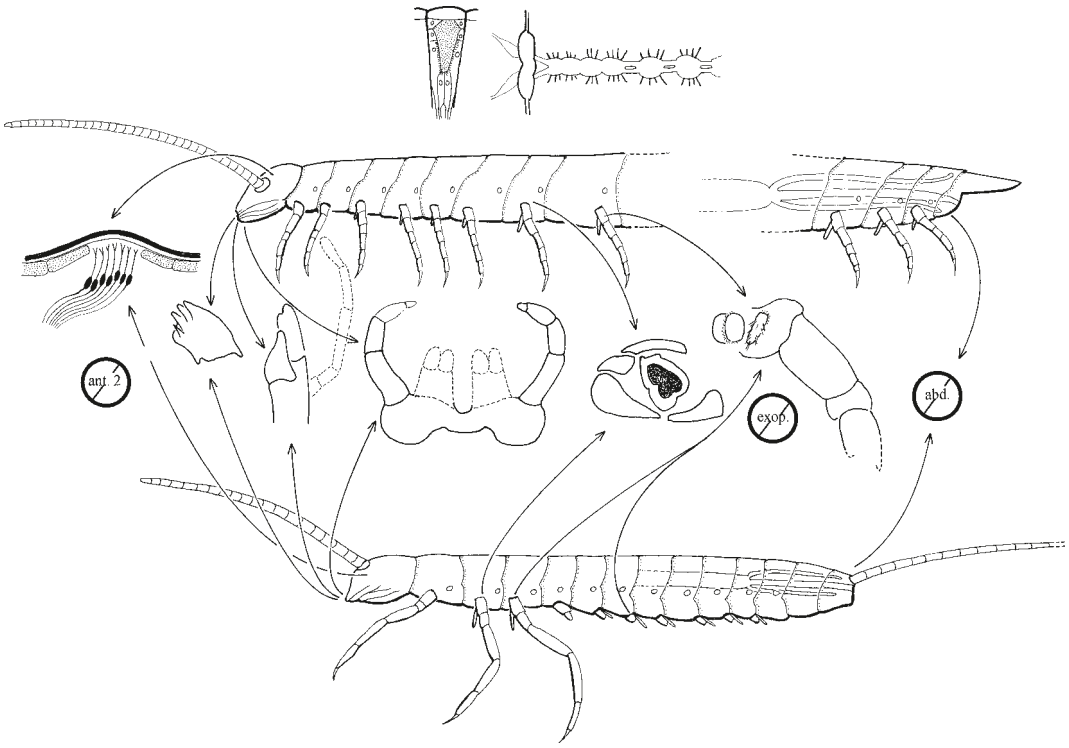


Fig. 88. Corresponding details in taxa of the Tracheata (top: Myriapoda; bottom: Insecta) together result in a complex pattern that increases the probability of homology for each detail of the ground pattern of Tracheata. Some details of disputed homology: tracheae and position of the spiracles, ommatidia with crystal cone (plesiomorphy shared with crustaceans), elevation of the brain, reduction of the second antenna (ant. 2), postantennal organ, mandible without palp, maxilla with 2 endites and without exopod, labium (second maxilla) basally fused and in the ground pattern with 2 endites and without exopod, existence of subcoxal sclerites, coxa with styli and coxal vesicles, walking legs with reduced exopods (exop.) (styli may be vestiges of exopods), reduction of the primary abdomen (abd., originally without traces of legs in other arthropods), ectodermal malpighian tubules.

evolve independently by chance. This is the **criterion of complexity** of homology research.

This criterion implies that with increasing complexity the probability of homology does not only increase for the whole pattern (the frame homology), but also for the individual details (modules) in it. The same detail has more weight in systematics when it is found within a complex and conserved frame homology than when it occurs in isolation (compare the "criterion of position", ch. 5.2.1). Single nucleotides in a specific position of a specific gene are informative, the isolated nucleotides have no phylogenetic value. In the analogy of Fig. 87, if two words of the same structure have a high probability of homology this is also true for the individual letter within these words. This is an example of the **principle of reciprocal illumination** which appears somewhat mystical without the probabilistic explanation: the details reinforce reciprocally their information content. An example from every day life: when we hear a single stroke of a key on a piano, or an isolated tone cut from a concert recording, we cannot guess to which piece of music the tone belongs. However, when we hear a few bars containing this tone, it is often possible to identify the piece of music (Fig. 89). With this finding not only the composition has been recognized, but also the origin of the single tone: with increasing probability of homology of a complex pattern (a piece of music) also the probability of homology for the detail (a tone) increases.

This connection has already been noted by Hennig (1950: 185): "Jede Eigenschaft der Holomor-

phe, jede Übereinstimmung und jeder Unterschied zwischen den Organismen wiegt also in der Phylogenetischen Systematik nicht absolut, sondern sie gewinnen ihr Gewicht, mit dem sie als Zeugen für den Grad der phylogenetischen Verwandtschaft auftreten, nur durch ihre Stellung im Gesamtgefüge der die Holomorphe des Organismus ausmachenden Einzeleigenschaften." ("Each property of the holomorph, each correspondence and each difference between organisms thus does not weigh absolutely in phylogenetic systematics, but they gain their weight with which they appear as witnesses for the degree of phylogenetic relationship only through their position in the whole composition of single properties of the organism's holomorph." Hennig defines the **holomorph** as the sum of all properties of a semaphoront; a **semaphoront** is a stage of life (larva, adult) of an individual.)

Note that complexity shared by two patterns also indicates that the probability of character change is low in relation to the time since separation from the last common ancestor pattern. Low process probability implies high probability of homology of shared details.

The same principle can also be phrased as **criterion of compatibility**: the larger the number of potentially homologous individual characters shared in a group of organisms, i.e. characters which are compatible in the sense that they fit **to the same ground pattern**, the larger is the probability of homology of the individual character. The probabilistic basis is the same as for the criterion of complexity. The Tracheata, for example, have characters for which the possibility of convergence has not been ruled out notwithstanding structural correspondence (malpighian tubules, tracheae), because these are adaptations to life on land. Together with more specific characters of the Tracheata (structure of mouth parts, of thoracal legs, postantennal organs, reduction of the second antenna, presence of subcoxal sclerites, lack of midgut glands, direct development: see Fig. 88), which are probably apomorphic homologies of the Tracheata, the probability of homology for tracheae and malpighian tubules increases if seen as part of the ground pattern of the Tracheata. Compatibility means in this case that a hypothesis of homology for an apomorphy fits to a second hypothesis of apomorphy, because both support the same hypothesis of monophyly, or, both are found in the same ground pattern (for reconstruction of ground patterns see 5.3.2). Ontologically it is the same to note that adding single synapomorphic characters ("elements of a ground pattern") a complex pattern of higher order can emerge.

In contrast to the criterion of complexity, which can be applied directly to material objects, the criterion of compatibility requires more assumptions, because it refers to a reconstructed ancestor: evidence has to be presented for each hypothesis of homology of characters in a ground pattern, the individual assumptions that details may be homologies are established prior to the reconstruction of the ground pattern. In addition, some hypotheses on the monophyly of subordinated taxa may be required before a ground pattern of a large group is reconstructed. Therefore, the risk that a complex ground pattern is based on false assumptions is higher than for a complex character that refers to real organs.

Attention: it is important to distinguish between the complexity of a real structure and the complexity of a ground pattern, which corresponds to the compatibility of hypotheses of homology (see above). A single material object (e.g., a skull of an individual cat) can be described without reference to any hypotheses of homology. The criterion of complexity is not needed. The comparison of different individual skulls of carnivores allows a statement on the homology of single bones or teeth, wherever individual details agree in their specific position and/or fine structure. The criterion of complexity is applied, either at the level of the skull (using the larger pattern as frame homology) or at the level of a single bone or tooth (using this as a frame homology composed of characteristic details). However, when the skull of "the Edentata" is compared to the skull of "the Marsupialia", one takes it for granted that the hypotheses of monophyly for Edentata and Marsupialia, respectively, are well founded and that a ground pattern of the corresponding skulls has been reconstructed correctly. For the comparison of the reconstructed ground patterns the criterion of complexity can be applied (in the variation of the criterion of compatibility). To describe homologies shared by taxa of the Tracheata (Fig. 88, insects and myriapods), the ground patterns of insects and of taxa of Myriapoda have to be compared to gain a state-

ment on the number of corresponding details which occur in these ground patterns.

The criterion of compatibility differs from the **criterion of congruence** of phenetic cladistics (compare ch. 6.1): putative novelties are congruent when mapping them in a shortest topology (see maximum parsimony method: ch. 6.1.2) they occur only once on the same stem lineage. In this case one can assume that they evolved only once, and therefore they must be homologous wherever they are found. In contrast to the criterion of compatibility, the criterion of congruence requires the reconstruction of a most parsimonious dendrogram as a first step, of course based on a data matrix. Statements on congruence depend therefore on (1) the selection, (2) the number, and (3) weighting of characters (because these determine the tree topology, and thus probability of homology is already implied in weighting), and (4) on the algorithms used to construct the tree (see also chapters 6.1.2, 6.1.10, Fig. 139). And therefore this criterion is burdened with many more assumptions than the criterion of complexity for material structures or the criterion of compatibility for characters in ground pattern.

The criterion of compatibility as well as the criterion of congruence both require the assumption that the characters belong to a ground pattern. In the first case, however, a pattern composed of hypotheses of individual ground pattern characters which belong to a single hypothesis of monophyly is evaluated (e.g., the Tracheata concept in Fig. 88) without reference to a complete tree. In the second case, the number of putative novelties found along an edge of a complete dendrogram that represents a dataset is the foundation for a congruence statement.

The criterion of congruence does not allow a "reciprocal illumination" of frame and detail homologies, because the analysis is focused on the number of character changes (or changes of details = character states) along an edge of a given topology. However, the latter is not the same as a frame homology. Note: the criterion of congruence on principle does only allow statements on the homology of character states (detail homologies, potential apomorphies), while the identity of the frame homology (= the positional homology) is considered to be background knowledge that is not tested with this criterion.

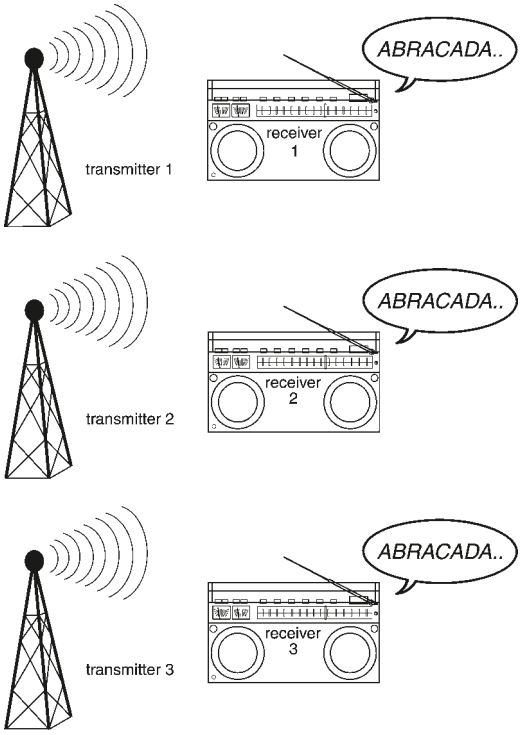


Fig. 89. The discovery that a common source of information exists corresponds to the recognition of a homology relation. Here three radios are depicted which receive signals of only one of many different radio stations. This circumstance can be realized without any knowledge about existing transmitters.

The compatibility method or tree construction by clique analysis (Eastabrook et al. 1977) is something else, it serves the grouping of taxa (see ch. 14.5). With the clique-method one can search for those characters which represent the majority of mutually compatible characters. This method is not suited for the reconstruction of dendrograms, because it does not allow an estimation of the quality of the dataset (see ch. 9.1).

Summarizing the previous reflections we can distinguish three levels at which the complexity of patterns can be evaluated. These levels show an **increasing uncertainty** in the sequence listed below, because the number of assumptions required as background knowledge increases:

– Level of the material object (**criterion of complexity** in the strict sense): only the assumption that we can trust our sense organs is required.

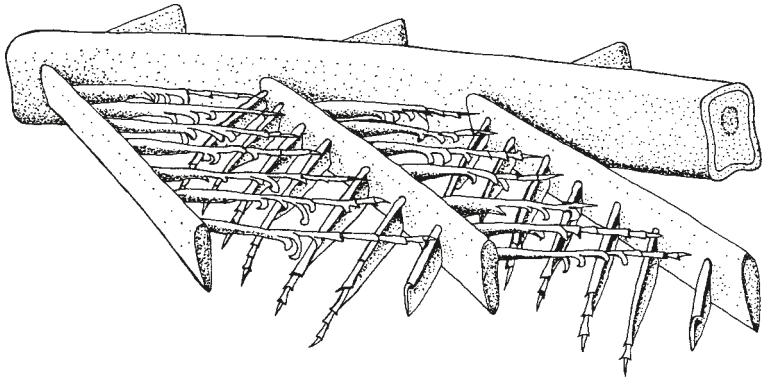


Fig. 90. Structure of a feather of a bird. The probability that this complex pattern evolved twice independently in nature is very low.

- Level of ground patterns (**criterion of compatibility**): additional assumptions are required on the homology and polarity of characters assumed to be present in the ground pattern, as well as the assumption of monophyly of the taxa that are being compared (e.g., Insecta and Myriapoda in Fig. 88).
- Level of the dendrograms (**criterion of congruence**): additional assumptions are that the method of tree reconstruction is realistic and adequate for the data at hand, and that the selection of terminal taxa and characters is representative for the real phylogeny.

These considerations on the evaluation of identities by no means concern the question how probable it is that a single character *evolves*: although the criterion of complexity is based on assumptions about the process of pattern evolution (Fig. 87), we do not analyse the process itself, but compare the end products of the historical processes to find out whether we are able to recognize that a common cause exists.

This approach is familiar to us from everyday life and comparable to a decision we make every day: when we hear for only a few seconds the same simultaneous sounds from two radios, we intuitively assume that both radios receive the same station. We know from experience, and maybe due to an inborn ability (see Riedl 1992), that the criterion of complexity (Fig. 87) is of importance in the real world. The probability that a complex pattern (sentence, melody, painting, sequence) develops twice only by chance is very small, and we arrive at a decision without analysing the process which produced these patterns. It is not relevant whether the simultaneously perceived melody has been broadcasted with a specific frequency range, if it was transmitted from a record or from a microphone, whether a piece of music is popular at the moment and has been sent for this reason. We evaluate the phenomenon, not the process that produced it. The same holds for the phenomenological analysis of characters.

We can make statements on the relative probability for competing hypotheses of homology stating which of the alternatives are supported by better (more) information. The extent to which details shared between two patterns (characters) are congruent determines the decision process. The more information is present, the greater is the probability that the decision is "correct", meaning that a real homology may have been identified. When only a few details are discernible this does not necessarily mean that these cannot be homologous, but we have to admit that our certainty for their correct identification is lower. This is the basis for the distinction between "valuable" or "good" and "weak" characters which is familiar to systematists. The experienced systematist will use for hypothesis of relationships only characters for which he or she assumes that they have a high probability of homology (Fig. 90).

The fact that the probability of a multiple evolution of identical complex structures is low does not mean that complex organisms should not evolve at all. The statement only means that a

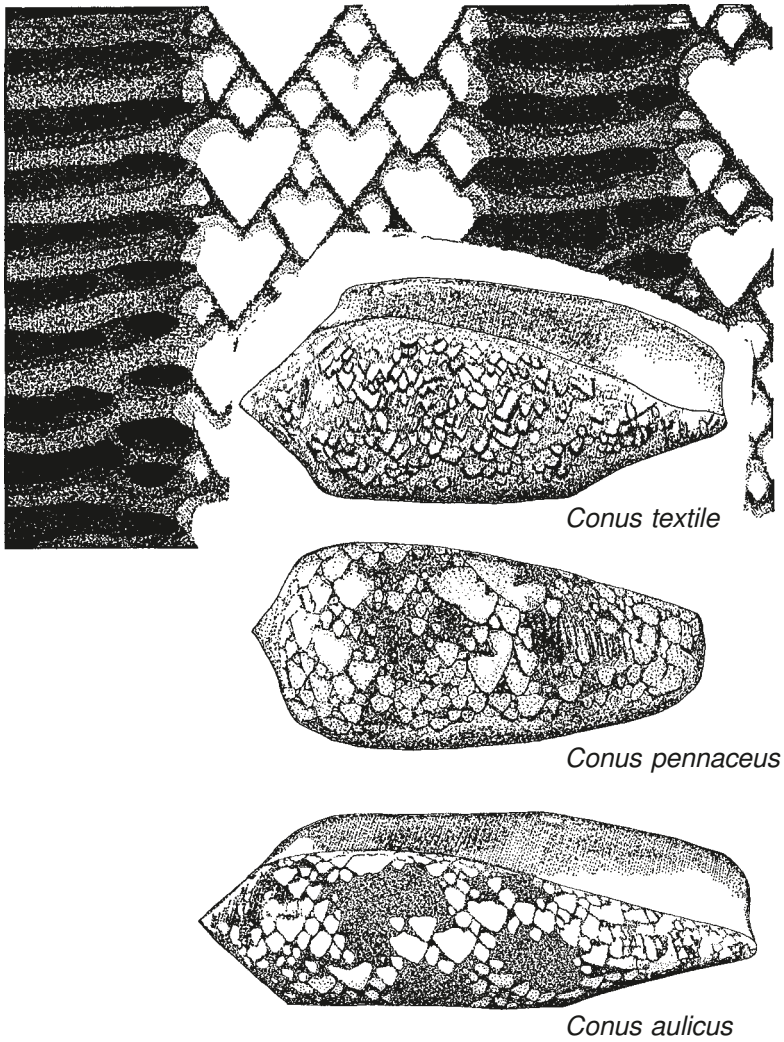


Fig. 91. What appears to be complex at first sight does not always have to be complex at the genetic level: colouration patterns of shells of related species of *Conus*, in the background a corresponding pattern produced with a mathematically simple model (after Meinhardt 1996, 1997).

second evolution of life forms (for example on another planet) would produce with greater probability organisms that differ in details of their construction from the ones found on earth. We can observe this rule even on our planet when convergent life forms evolve independently (compare for example vultures (Cathartidae and Accipitridae, Fig. 70), digging mammals (*Notoryctes*, *Chrysochloris*, *Talpa*, *Spalax*), eel-like fishes among Anguilliformes, Mormyriformes, Siluriformes, Dipnoi, succulent plants (e.g., *Adenium*, *Dorstenia*, *Chorisia*, *Dendrosicyos*)).

The criterion of complexity has been known for a long time. It is the estimation of the probability of co-occurrence of congruent but independent characters that was already recommended for phylogeny inference by K. Lorenz (1943). W. Hennig (1950: 175) talks about the "criterion of the complication ("Kompliziertheit") of characters".

The **detection of the complexity** of morphological novelties and of their uniqueness is a central problem systematists have to solve. They have to distinguish epigenetic variations from inherited

ones. An apparently complex variation of a morphological character must be the phenotypic expression of a novelty at the level of genes. Similarly, novelties in a DNA region could form a complex homology. However, whereas the changes of a DNA sequence can be recorded quantitatively, morphological data that allow a quantification in relation to genetic changes are usually not available. One has to rely on indirect clues.

Examples: the pigmentation patterns in shells of *Conus* species (Mollusca: Gastropoda) are produced by the activity of pigment-synthesizing glands in the zone of the mantle that secretes the shell's margin. Pigment free areas probably originate due to a reciprocal inhibition of pigment production by neighbouring groups of cells and due to the varying concentration of raw material needed for pigment synthesis (Fig. 91). Mathematical models of these interdependences allow an artificial creation of these patterns. Small variations of model parameters result in the formation of apparently complicated new patterns similar to those occurring in nature in different *Conus* species. Whether in nature also only few mutations are sufficient to produce these variations of pigment patterns can only be inferred indirectly, because the genetic basis is not known: the comparison of different *Conus* species shows that essential elements of shell patterns are conserved, whereas number, size and position of elements show a higher variability which also occurs intraspecifically. This allows the conclusion that the genetic basis for the variation has to be simple, single variants cannot be complex novelties. Similar arguments are true for variations in fur colouration of mammals (e.g., black patches), for the number of cuticular hairs in bristle fields of arthropods, proportions of skull bones in humans, or for variations in wing colouration of butterflies.

Defect mutations and **reductions** (also deletions, "negative characters") are usually based on only few mutations. Furthermore, different mutations can produce the same phenotype, which may therefore occur several times convergently. Defect mutations normally are not complex characters, and for such novelties a low probability of homology has to be assumed (see Fig. 94, convergent reduction of eyes).

Examples: the search for the genetic basis of defects causing diseases is financed more easily for humans than for other creatures. It is known that the phenotypically visible malfunction known as *Dystonia musculorum deformans* (irregularity of movements) can partially be ascribed to the deficiency of Dopa (= Dihydroxyphenylalanine). This deficiency can be caused by very different mutations: among others, the enzyme TyrH and the cofactor BH_4 are involved in the synthesis of DOPA. The synthesis of the latter again depends on the enzyme GTP-cyclohydrolase I. Single point mutations in different parts of the genes coding for these proteins cause the same phenotype, therefore its probability of homology is low (e.g., Ichinose et al. 1994). – The evolution of resistance against certain insecticides in insects is a similar unspecific character. This can be deduced from the fact that resistance can evolve quickly and many times independently (resistance has been documented for about 500 species of arthropods). The molecular causes can be, for example, point mutations in genes of the acetylcholinesterase or of GABA-receptors (Alzogaray 1998).

Discussing the quality of characters often the **criterion of independence** is stressed to be important. However, this statement is frequently based on erroneous arguments. It implies that two functionally dependent characters do not have the same value for phylogenetic analyses as two independent ones. Dependence here means that the presence of some detail also inevitably causes the presence of another detail. For example, when a stridulating insect has a dentated ridge, a complementary piece that brushes over the ridge also has to be present. The development of wings requires the development of appropriate muscles. A nucleotide substitution in a helical region of a RNA molecule requires a complementary substitution in the complementary strand to conserve the secondary structure (Fig. 92).

First of all the question has to be raised why independent consideration or weighting of functionally linked characters should be avoided. To understand this problem, genetic and functional dependence have to be distinguished.

Genetic coupling: for weighting of characters only the probability of homology is relevant. We have to ask: in which cases is the probability of homology reduced by the functional dependence

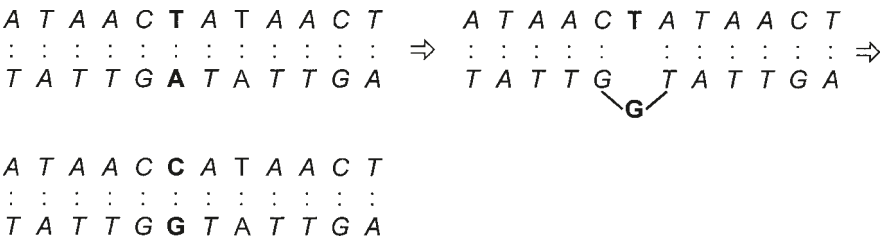


Fig. 92. Dependent substitutions: the second substitution event compensates the consequences of the first mutation, but it is an independent evolutionary novelty.

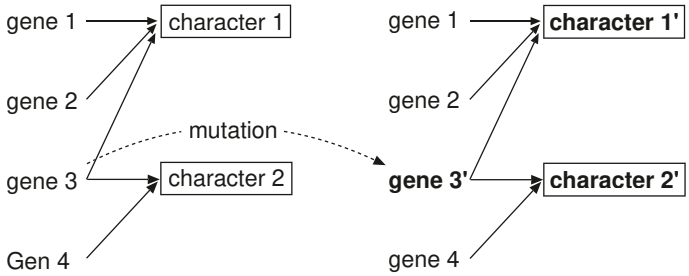


Fig. 93. Dependence of morphological characters: when one gene influences several characters ("polypheny" or "pleiotropy") a single mutation can modify two (or more) characters.

of two details which could be homologized separately? Obviously only when the presence of two (or more) novelties **feigns** the occurrence of two (or more) **events**, where only one event took place, for example when a mutation in a single gene causes modifications in two or more characters (scheme in Fig. 93). Adding the number of visible corresponding novelties of two organisms, the number of coded potentially homologous mutations is higher than in reality.

Examples: pleiotropies are well known from intraspecific mutants: alleles which determine the colour of flowers can also influence the colour of seeds and leaves. Mutations in mice influence at the same time fur colouration as well as bone growth. In humans the combination of "spindle fingers" and defects of the eye lens is known. – Many lizards have no legs (e.g., Anguinae) and thus toes are missing as well. Should it be shown that a gene which induces the transcription of further genes necessary for leg development is inactivated in a legless species, the absence of all leg characters could possibly be the effect of a single event. However, in case it can be shown that the reduction of appendage characters occurred step by step, each event can be counted individually.

Functional coupling of "positive characters", when structural details are exchanged or added, must be regarded in a different way. The lens eye of an octopus (Cephalopoda: Octopoda) is more complex than the pinhole eye of a more primitive nautilus (Cephalopoda: Nautilida). The eyes of different octopodids share many detail homologies absent in *Nautilus* and it is justified to count these details individually, although they are functionally coupled. In the case of complementary substitutions in DNA sequences (Fig. 92) the second mutation is advantageous for functional reasons, because it restores the original secondary structure of the molecule. When a modification of the secondary structure is under selection pressure, the probability of events is higher in a population for a substitution that complements a first mutation than for the conservation of the older state. This is, however, irrelevant for the phenomenological character analysis and for weighting of characters. The second mutation will not occur inevitably, and when two organisms show both mutations, more information is available than when only one mutation can be found. Therefore in a phenomenological approach it is correct to count both mutations separately to evaluate the "probability of cognition" (see ch. 1.4.5). In many cases the functional coupling of

nucleotides or amino acids in sequences (e.g., via tertiary structure) is not known. The sequences can nevertheless be used for phylogenetic analyses. The consideration of probabilities of evolutionary events is only convenient when there are good reasons for the assumption that this probability is estimated correctly. This argument holds for morphological as well as molecular characters.

The same is true for physiological and behavioural characters. For example, it is to be expected that lizards search actively for places where the temperature is optimal for their physiology. Temperature preference depends on the adaptation of the musculature to the average temperature of the environment. This probably requires a coadaptive modification of several genes. Analysis of Australian skinks (Lygosaminae) has shown that temperature preference evolves faster than the adaptation of the locomotory apparatus (Huey & Bennett 1987). Coevolving characters can therefore be counted as independent characters. – In frogs many sexually dimorphic characters are under control of male hormones (androgens). Parallel reduction of hormone production would influence nuptial pads, vocalization, forearm flexor size, etc. Applying lower weights to these correlated characters did not produced a better resolved phylogenetic tree than when considering each character separately (Emerson 1998).

All functional characters of an organism depend on each other. Even when morphological structures are not functionally coupled in an obvious way they nevertheless depend on the existence of the other ones whenever they are important for the survival of the organism. Hairs and skeletal musculature may be regarded as being functionally independent. However, when the animal freezes to death or suffers from a muscular disease all other characters are not transmitted to the next generation. – Probably only characters which are neutral to selection forces are functionally independent. These, however, have from the point of view of a systematist the big disadvantage to be too variable, a result of the lack of selection (e.g., shape of the human earlobe, absolute number of hairs, absolute number of chromatophores, position of folds in the skin, etc.). Their variability renders these characters useless for analyses of interspecific phylogenetic relationships, they are too "noisy". They may sometimes be informative for intraspecific characterization of groups of individuals.

Hypotheses of homology are testable. They allow predictions on phylogenetic hypotheses, which are either verified or rejected with further characters (s. ch. 1.4.3).

5.1.2 Weighting

Weighting of characters serves the differentiation of their estimated probabilities of homology. Even the (conscious or unconscious) selection of characters is a form of **weighting**:

	Character is considered	character is not considered
Weight:	1	0

In cladistic analyses it is possible to weigh in a very differentiated way, for example, to assign characters weights between 1 and 20. This presupposes that relative ranks of probabilities of homology can be distinguished. Even when such a differentiation is possible, the assignment of discrete numbers representing probabilities is a very subjective decision. In cladistics a weight can methodologically also be considered in form of "**costs**". We could assume that a character of high complexity requires high energy costs for its evolution. However, this is not the significance of "costs" in cladistics. The term "costs" is used to count character state changes for the maximum parsimony method (Fig. 123). In a dendrogram, the "expensive" character has the same value as several simpler characters (see ch. 6.1 and Fig. 129). Therefore, the "cost" represents the estimated probability that a character is a homology.

The term "costs" used in cladistic literature could refer to a) energy requirements during evolution or b) the methodological consequences of weighting. Concerning a): how many mutations are necessary to construct a new character, how many individuals have to conclude their life cycle and produce offspring to guarantee the dissemination of the new character in a population, and how many malformations are selected is unknown. Such costs cannot be estimated for historical processes. But, it can be assumed in this

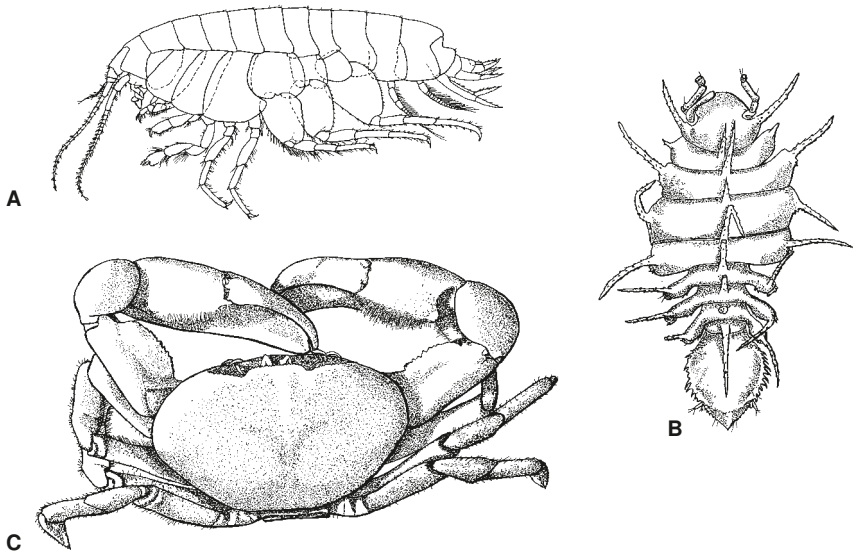


Fig. 94. Reduction of eyes in deep sea crustaceans. **A.** *Venetiella sulfuris* (Amphipoda: Lysianassidae), **B.** *Notoxenoides dentata* (Isopoda: Paramunnidae), **C.** *Austinograea williamsi* (Brachyura: Bythograeidae) (after Barnard & Ingram 1990, Hessler & Martin 1989, Menzies & George 1972). The frequent reduction in different taxonomic groups indicates that eye reduction is not a very complex character. Therefore, probability of homology is estimated to be low.

sense that complex characters cause higher costs. Concerning b): a higher weighted character increases the "length" of a tree. These "costs" are chosen by a systematist or by an algorithm and have no measurable relation to the real energy and time expenditure during evolution. In this sense, when using the MP-method "costs" are added to the "overall length" of a topology ("maximum parsimony algorithms", see ch. 6.1.2).

Theoretically, weighting could concern two different aspects:

a) the probability that a group of evolutionary novelties **evolves and is retained** (weighting according to an evaluation of the probability of events), and
b) the probability, that a real homology **can be identified correctly** (in ch. 1.4.5 called "probability of cognition").

An example for weighting of probability of events is differential weighting of transitions and transversions (see ch. 6.3.1). When a transversion is counted twice this does not mean that the character is more complex, but that it is retained in a population with a higher probability.

In the practice of comparative morphology, considerations about the course of evolution as well as on the evaluation of the complexity of visible structures are taken into account. However, the "probability of events" can generally not be quantified. It makes therefore no sense to use model-dependent methods of phylogenetic analysis (ch. 8). Example: in deep sea animals which evolved from shallow water species with eyes, often the eyes are reduced (Fig. 94). The probability that a hypothesis of homology is well founded can be estimated in two ways (see ch. 4.3.1):

A) **Evaluation of the probability of events**: it can be assumed that a loss of vision can develop through different mutations and that probably often only few mutations are necessary. In the deep sea, these mutations have less disadvantages compared to the same mutations in shallow seas or they are beneficial due to the saving of material and energy. Thus blindness can evolve often convergently, the character "blindness" has a small probability of homology and thus gets an arbitrarily chosen low weight in cladistics. However, the evolutionary process is not known for the individual case. We explain the assumed proc-

character	number of characters
tympanic cavity with three ear ossicles	1

character	number of characters
malleus present	1
manubrium mallei attached to tympanic membrane	1
incus present	1
columella shortened to a *stapes*	1
	sum: 4

Fig. 95. Partitioning of complex morphological characters leads to a higher weight.

esses with analogous cases (e.g., with the observed frequency of defect mutants in animal breeding). – Bones in a skull evolve with covaration, the probability of change in one bone can depend on that of another bone. Inference of possible relationships could be used in a weighting scheme.

B) **Evaluation of the probability of cognition**: we do not know whether in closely related deep sea species eyes were already missing in the last common ancestor and how probable it is that the required mutations can occur in the time available along the stem lineage of a taxon. Therefore we choose other arguments for weighting: the character "eyes missing" (without further specifications) is not complex enough to establish an important hypothesis of homology, although there is a chance that it is an evolutionary novelty and thus an (apomorphic) homology of the species in question. The lack of complexity justifies the character's low weight, it corresponds to the uncertainty for the identification of a homology.

"Weak" characters cannot substantiate a sistergroup relationship. However, sometimes they may be interesting for evolutionary biology and may "fit into the scenario". They can be compatible with a well-founded hypothesis on phylogeny and be congruent with a hypothesis on the evolution of ways of life. When a group of organisms comprising exclusively deep sea animals is monophyletic, the character "eyes reduced" is compatible with the hypothesis of a radiation in the deep sea starting with possibly blind deep sea ancestors. This "fitting" increases the probability that the character is a real homology (either due to the criterion of character compatibility in a ground pattern and/or due to parsimony in a topology). Note, that in this case the tree must be constructed first before the weak character is used in the argumentation.

The "weakness" of characters, in other words, their low probability of homology, can only be assessed in the context of the complete data matrix. Weak characters are usually fast evolving and therefore too noisy to infer ancient divergence events, but they may have a different importance in an analysis of closely related species. Distantly related species will show with higher probability chance similarities, but in a dataset of similar species (or of individuals within populations) the same character states can be highly informative.

As explained before, **weighting of morphological characters** has to be based in practice on the estimation of the probability of homology. A practicable method consists in mentally partitioning a complex character into its parts and to list these separately in a data matrix, provided that these details are thought to be also with high probability homologies. In this way, a higher weight results that depends on the detection of complexity. For the mammalian character "tympanic cavity with three ear ossicles" total weight can be increased as in Fig. 95 and, knowing the details, even to a higher value. Through further analysis of the fine structure of the ossicles and associated structures as the innervation and the musculature, this list could even be elongated. Doing this the structural complexity is accentuated in such a way that it becomes visible that the pattern "ear ossicles" is a unique peculiarity of mammals. This requirement to examine characters closely also has the advantage that a superficial similarity is recognized as such and is not included in a data matrix.

Probability of homology: either the estimated probability that a real homology has been recognized correctly or that a homology evolved.

Weighting: differentiation of characters according to a ranking of probabilities of homology, evaluated according to the probability of cognition or of the probability of events.

Criterion of complexity: the greater the number of corresponding elements in two patterns the higher is the probability that these patterns have been produced by the same process, or rather have the same source of information. In systematics this criterion refers to the properties of material individuals and serves the evaluation of probability of homology within a group of organisms.

Criterion of compatibility: the larger the number of characters occurring in the same group of taxa and matching in fine structure and/or position, the greater is the probability that the individual character is homologous among the taxa of the group and part of the ground pattern of the group in relation to other such groups. This criterion refers to ground patterns and serves the evaluation of probability of homology and, in case of evolutionary novelties, the estimation of probability of monophyly of a group.

Criterion of congruence: characters are congruent when they appear as potential evolutionary novelties on the same edge of a given topology. This criterion depends on the fit between a data matrix and a given topology. This is the sole criterion of homology in *phenetic cladistics* (see 6.1).

Attention: While weighting morphological characters in cladistics (ch. 6.1.3), it has to be distinguished whether a frame homology or a detail homology is being evaluated. It is not correct to weigh the frame character (equivalent to a single column of the matrix) as a single entity, because the probability of character state changes within this frame may differ for each state. The probability of homology is greater for the complex frame than for the detail. Due to a confusion of frame and details, character state changes may get an exaggerated weight.

Weighting of **sequence data** (ch. 5.2.2.2) according to pattern complexity is only convenient for phenomenological approaches (ch. 6). Using model-dependent methods (ch. 8), the probability that a specific process occurred is assessed, but not the probability of homology of single character states of terminal taxa.

5.2 The search for morphological and molecular homologies

In the following paragraphs no laboratory protocols for histological or molecular analyses are presented, these can be found in laboratory manuals. We will examine the question, which of the data gained with laboratory methods are valuable for phylogenetic analyses and focus on the assessment of the probability of homology of characters.

5.2.1 Criteria of homology for morphological characters

In the preceding chapter a theoretical foundation has been presented which justifies an objective assessment of alternative hypotheses of homology. For practical work there exist several criteria that have been proposed to identify homologies. A careful consideration of these criteria shows that they can only be used without contradictions when the term homology is defined as in ch. 4.2 (“identity of inherited information”).

Remane (1952, 1961) proposed three criteria of homology for morphological characters (criterion of position, of special quality, of continuity) all of which can be traced back to the criterion of complexity. Further criteria can be found scattered in the more recent literature.

The **criterion of position** is the assumption that a detail can be homologized when it always has the same position in relation to neighbouring details of a complex pattern. In sequences of macromolecules, we talk of positional homology when a specific *sequence position* or a longer insertion is homologized with alignment techniques (ch. 5.2.2.1). Alignment of a sequence position is only possible when the neighbouring regions are conserved and thus are recognizable in different

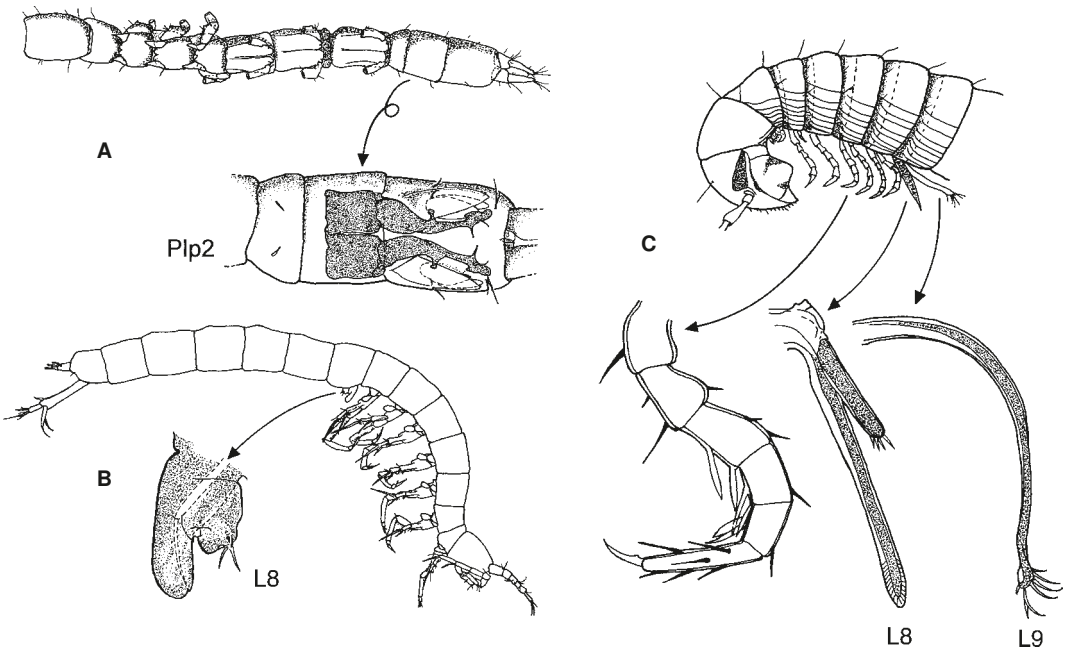


Fig. 96. Highly modified male gonopods can be homologized with normal legs due to their position. **A.** Modified pleopod at the second pleonite in microcerberids (Crustacea: Isopoda). **B.** Modified eighth leg in bathynellaceans (Crustacea: Syncarida). **C.** Modified eighth and ninth legs in diplopds (after Schminke 1987, Schubart 1934, Wägele 1982).

organisms (compare Figs. 98, 103, 106, 149). The more complex or specific the surrounding patterns are, the more successful and unambiguous the homologization will be.

We can state that the mandible of different Crustacea is homologous regardless whether it is a cutting or chewing tool, a stiletto (as in parasitic larvae of the Gnathiidae (Isopoda)) or a reduced appendage bud without function (e.g., in sexually mature Gnathiidae or in many adult Sphaeromatidae that do not feed): the relative position to other parts of the head (between labrum and hypopharynx, anterior to maxilla 1) allows the identification. The same argument can be used for serially repeated but modified organs of an organism. Gonopods, for example, can be found in many arthropods at a position where other body segments or segments of immature individuals bear normal legs (Fig. 96), wherefore gonopods are homologized with these appendages in the sense of a secondarily modified leg.

However, an important question remains: which of the many details are the constituent parts of a homology, or which is the shared inherited information that is indirectly visible in the phenotype? In the case of gonopods it is obvious that there exists new information that was absent in a last common ancestor (the one with normal legs) and which is responsible for the species-specific modification and specialization of gonopods in individual taxa and species (compare among crustaceans the modified pleopods of male decapods or peracarids, sixth thoracopods in copepods, eighth thoracopods in Bathynellacea). These taxon-specific novelties cannot be homologized with older information which codes for the ontogenetic development of normal thoracic legs. But the position of the anlage of an appendage within the fabric of segmental sclerites and muscles and nerves remains the same compared to unmodified legs. Since many details in the fine structure of legs and gonopods are not homologous, there is at least the homology of the anlage at a specific position and the same contact of the appendage with neighbouring structures. Positional homology is often the result of identity of developmental mechanisms.

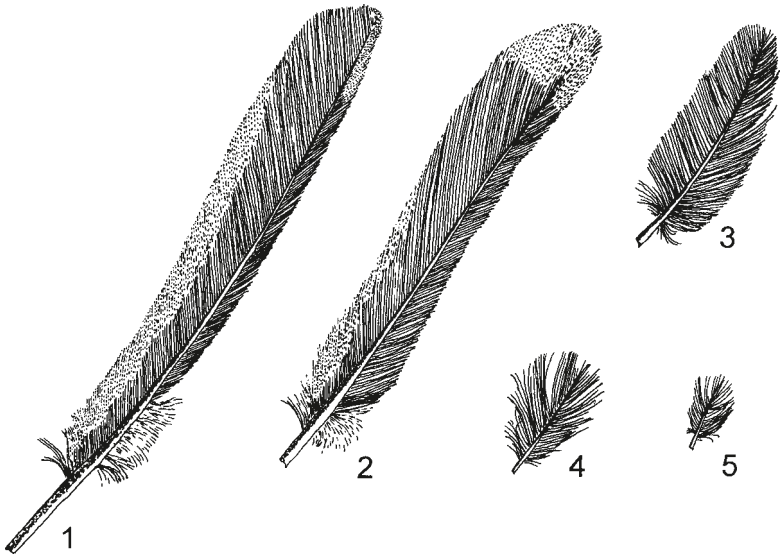


Fig. 97. Feathers of birds are homologous things. Homology means in this case that there must have existed an ancestral feather with specific characters that are still present today (characters of the ground pattern of the feather). However, comparing different feathers of the same animal, many differences can be seen: these are based on genetic information which is most likely not identical for all feathers and thus not homologous. When feathers are compared, the constant details found in all feathers belong to the frame homology, while variable details do not all have to be homologous and do not belong to the constant frame. Order of the depicted types of feathers: secondary, dorsal secondary tectrices, axillary feathers, dorsal primaries.

When the criterion of position is not fulfilled, characters may nevertheless be homologous. Just think of the different position of feathers, scales or hairs, of plastids, mitochondria etc., which are constructed identically independent of their position. Let us consider the case of the *antennapedia* mutant of *Drosophila melanogaster*, which develops a thoracic leg on its head instead of the antenna: here apparently the criterion of position fails. The appendage inserts in a position where other individuals of *D. melanogaster* carry an antenna. Nevertheless, the leg of the mutant cannot be called an antenna, because it has the fine structure of a leg. This example demonstrates that the existence of specific contacts between a structure and its surroundings are part of the criterion of position. In insects, the contacts and spatial relations between the coxa of the thoracic legs and the neighbouring pleural sclerites are homologous, but these contacts do not exist in the case of the head leg of the mutant. Therefore the point of insertion of the head appendage of the mutant is not homologous to the one of thoracic appendages. But the leg segments and joints can well be homologized between the leg of the head and thoracic legs, because even in the "wrongly" positioned leg, the tibia, for example, is found between femur and tarsus and has the characteristic joints, hair sensilla, etc. It shows besides the "correct" position within the frame "thoracic leg" also the same "specific quality". It is therefore also a mistake to deduce that antenna and leg are homologous because they can develop in the same part of the body! These phenomena are easily understood when we remember that homologous genes can be activated in different regions of a body.

Obviously, in morphology the criterion of position refers to identical genetic information which induces in the course of ontogeny the anlage of an organ. With increasing knowledge of molecular developmental biology it will become possible to identify this homology at the level of genes. In comparative morphology, the probability that such a homology is present increases with the number of specific elements (e.g., specific sclerites, bones, nerves) that show a constant position relative to each other. In principle, with the criterion of position a complex pattern is evaluated and not only a single detail.

The **criterion of specific quality** is also a consideration of the complexity of patterns. When sim-

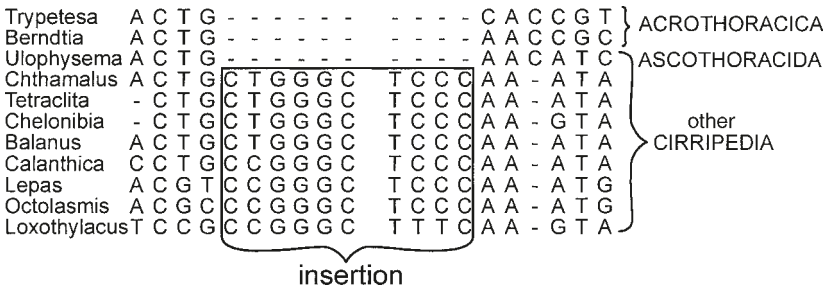


Fig. 98. The criterion of specific quality can also be applied to molecular characters: identical construction of an insertion of the 18S rRNA-gene of crustaceans of the taxon Cirripedia.

ilar characters found in different organisms are placed in dissimilar surroundings, a hypothesis of homology is only well supported when there are corresponding details within these characters. Feathers (Fig. 97), for example, are always homologous in the sense that there are specific genes which carry information for the construction of a dermal papilla, a keratin sheath and in it the supporting calamus and the shaft, independent of whether the genes are activated on the wing or elsewhere. There are, however, differences in size, microstructure and colour between contour and downy feathers, for example. These differences also must have a genetic basis. Even though the cascades of gene activations leading to a specific morphogenesis are not completely known, the variations of the morphology of feathers furnish evidence for the simultaneous presence of different genes, of which only some are activated for the construction of one type of feather.

In DNA or protein sequences the specific quality is a specific series of nucleotides or amino acids. When the same "signature" consisting of many elements occurs repeatedly in different organisms, a hypothesis of homology is with higher probability correct than a hypothesis of analogy. An insertion (Fig. 98) can be regarded as a single pattern and is homologized due to its specific quality.

The **criterion of continuity** is relevant to homologize patterns that contain only few congruent details. When two dissimilar patterns are linked by intermediate forms during ontogeny or phylogeny, showing that a pattern is a copy of another one, the patterns can be homologized in the sense of frame homologies. This corresponds to the identification of the same detail in a series (as in Fig. 77), where the starting state can be compared with the terminal state using intermediate forms.

At a closer look, the criterion of continuity is a specific application of the two preceding criteria. A homologization can be possible through the study of **ontogeny**, whereby for the most part the criteria of position and/or of specific quality are used to compare embryonic structures with the ones of adult organisms. The pair of ventral suckers on the head of parasitic "fish lice" (Branchiura, Crustacea) can be identified as first maxillae because ontogenetically they develop from appendage buds which in the larva are found in a position where usually the first maxilla of other crustaceans occurs. The argument relies on an observation of the position in the larva, followed by the direct observation of the further ontogenetic development of this structure. Another well known example is the homologization of the lower jaw bones of gnathostome fishes with ear ossicles (malleus and incus) of mammals (theory of Reichert-Gaupp). During embryogenesis of mammals, Meckel's cartilage, which corresponds to the primary lower jaw of primitive Gnathostomata, develops in the caudal part as the jaw's articulation, but is later transformed into the malleus. – Human embryos show externally visible pharyngeal arches in the region in which gills develop in fishes. The four embryonic branches of the aorta present in mammals can be homologized with gill vessels due to their position. A homologization with the aortic arches of other vertebrates is possible. In mammals they develop in such a way that only the second left vessel forms the aortic arch.

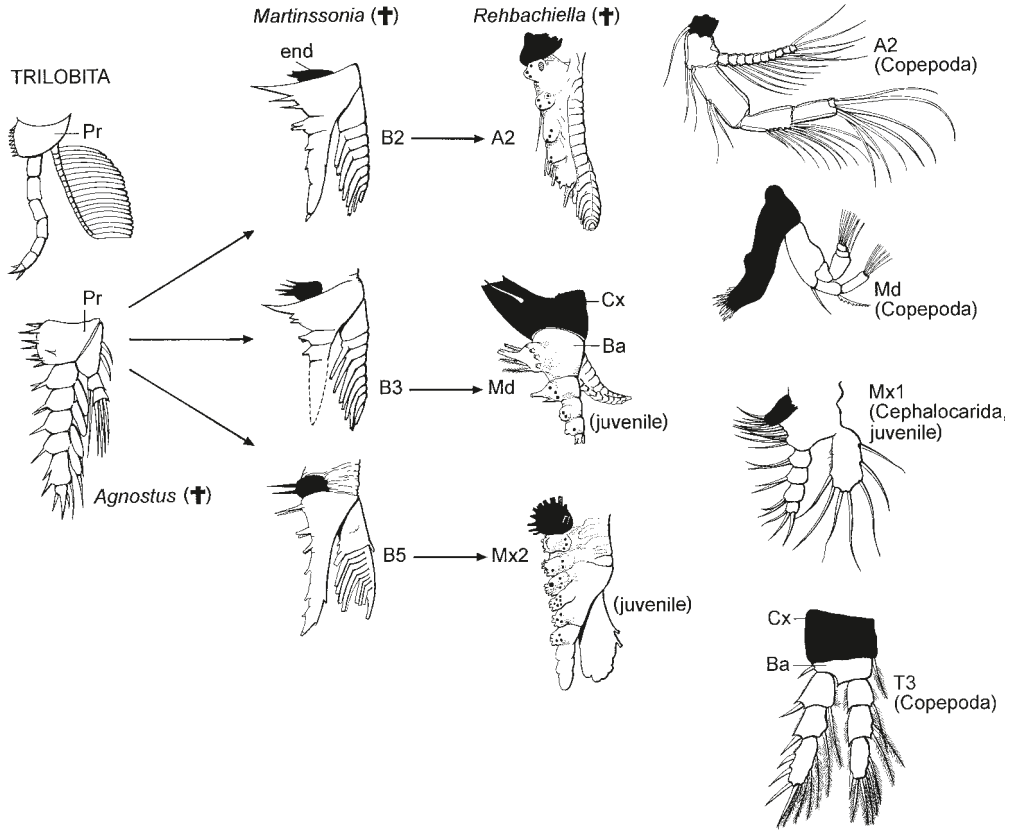


Fig. 99. Well preserved phosphatized fossils (e.g., *Agnostus*, *Martinssonia*, *Rehbachiella*) show the evolutionary development of the protopod of mouthparts and thoracic appendages in the stem lineage of the Mandibulata. The criterion of continuity can be applied here to homologize coxa (Cx) and basis (Ba) of the Crustacea with the protopod (Pr) of trilobites. The coxa develops from the proximal endite (end) of the protopod. The arrows do not indicate ancestor-descendant-relationships but the postulated changes of construction principles. A2: second antenna; B2-5: postantennal appendages; end: proximal endite; Md: mandible; Mx1, 2: first or second maxilla; Pr: protopod; T3: third thoracopod. (Modified after data in Walossek 1993 and further figures after Huys & Boxshall 1991, Müller & Walossek 1986, Schram 1986).

The criterion of ontogenetic origin **cannot be reversed**: as on principle (with few exceptions) each cell contains the complete gene inventory of an organism, organs, cellular structures and molecules can be homologous, although they develop during ontogeny from different anlagen, different tissues, or at different positions. It is not surprising that in vertebrates mesenchym as well as derivatives of the neural crest can produce cartilage (see Thorogood 1987). The site of production is not relevant when the homology of hairs or of erythrocytes of mammals is discussed. Statements of homology can also be applied to molecules: the lens proteins in the vertebrate eye (crystallines) are not only produced in the eye, but often also serve as enzymes in cellular metabolism (Wistow 1993). Knowing this, the argument that the halteres of Strepsiptera may be homologous to those of the Diptera although they occur on a different thoracic segment can be understood (Whiting et al. 1997; however, convincing molecular synapomorphies for a sistergroup relationship Diptera/Strepsiptera have not been found so far, Hwang et al. 1998).

Further applications of the criterion of continuity arise when **fossils** are known that can be regarded to be **intermediate forms**. The homologization of coxa and basis of the thoracic leg of higher Crustacea with the protopods of trilobites or

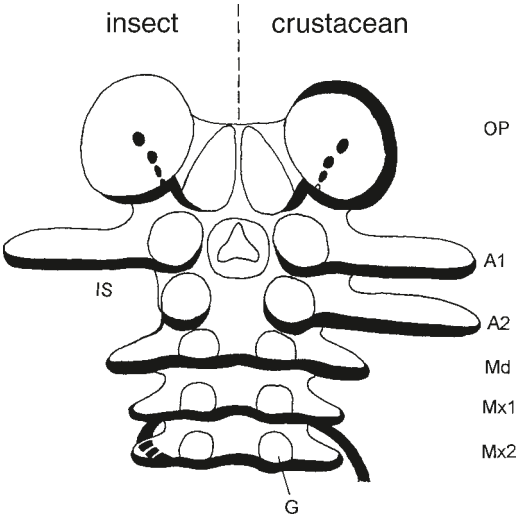


Fig. 100. Homologization of embryonic segments based on gene expression patterns. Schematic illustration of the *engrailed* expression in the head of insects (left) and crustaceans (right). The series of ganglia, leg anlagen and the *engrailed* expression prove that the intercalary segment (IS) of insects (and accordingly also of Myriapoda) corresponds to the segment of the second antenna of crustaceans. OP: optic-protocerebral region; A1, A2: first and second antennae; Md: mandible; Mx1, Mx2: first and second maxillae (modified after Scholtz 1997).

Chelicerata is considered established only since Cambrian fossils were discovered in which the stepwise separation of the coxa from the protopod can be seen (Fig. 99). This is also a specific case application of the criterion of position.

Summarizing, Remane's criteria of homology can be generalized to the criterion of **complexity** as explained in the preceding chapter. Further auxiliary criteria thought to be useful for practical work are known:

Criterion of the expression of homologous genes: the demonstration of the expression of segment polarity genes (e.g., *engrailed* or *wingless*) is used to postulate the presence of segment anlagen. The position and number of anlagen allows the homologization of individual segments, when in a series of metameric anlagen additional markers exist which can serve as point of reference to start segment counts (e.g., the anlage of specific, easily identified appendages). Problems may arise when single segments are completely reduced without trace (antennal segment of spiders?).

As long as genes are studied which are necessary for the development of a specific structure and which are exclusively expressed for its development, a homologization of this structure due to homologous gene expression is logically correct. It is, however, undue to use any gene expression to homologize structures of the phenotype. For example, hairs of the skin of mammals grow on very different parts of the body, but nobody would try to derive from the corresponding gene activities an argument to homologize different body parts (head and finger, for example). This mistake has occurred in less obvious cases. The homologous genes *Pax-6* (vertebrates) and *eyeless* (insects) are expressed early during embryonic development of eyes and induce cascades of transcriptions of other genes that finally end with the formation of an eye. This, however, does not mean that the compound eyes of insects are homologous to the lens eyes of vertebrates. The correct statement is that homologous gene products have some function during early embryonic development of eyes, other homologous genes may be active later: the *Pax*-6 gene is not the complete genetic information necessary for the construction of a vertebrate eye. The vertebrate lens eye evolved *de novo* to its present-day complexity independently of the insect eye, which had different precursors. However, for both types of eyes some (how many?) products of homologous genes are recruited, and possibly both eyes contain modified versions of the same ancestral type of simple sensory cell (which does not necessarily deserve the name "eye"). – The *distal-less* gene is expressed during the development of legs of arthropods and of tetrapods. A homology of the legs of these groups of animals cannot be deduced from this fact.

An example of this logical mistake is the theory that epipods of crustaceans are homologous to insect wings (Averof & Cohen 1997). The chain of reasoning of these authors is:

1) wings and thoracic legs of insects develop from the same embryonic regions.
2) Genes which are homologous to *apterous* (*ap*) and *pdrn* of *Drosophila* were found in a crustacean *(Artemia)*. (A homologous gene also occurs in vertebrates and echinoderms!).
3) In *Drosophila, pdm* and *ap* are expressed during early embryonic development of wing and leg buds,

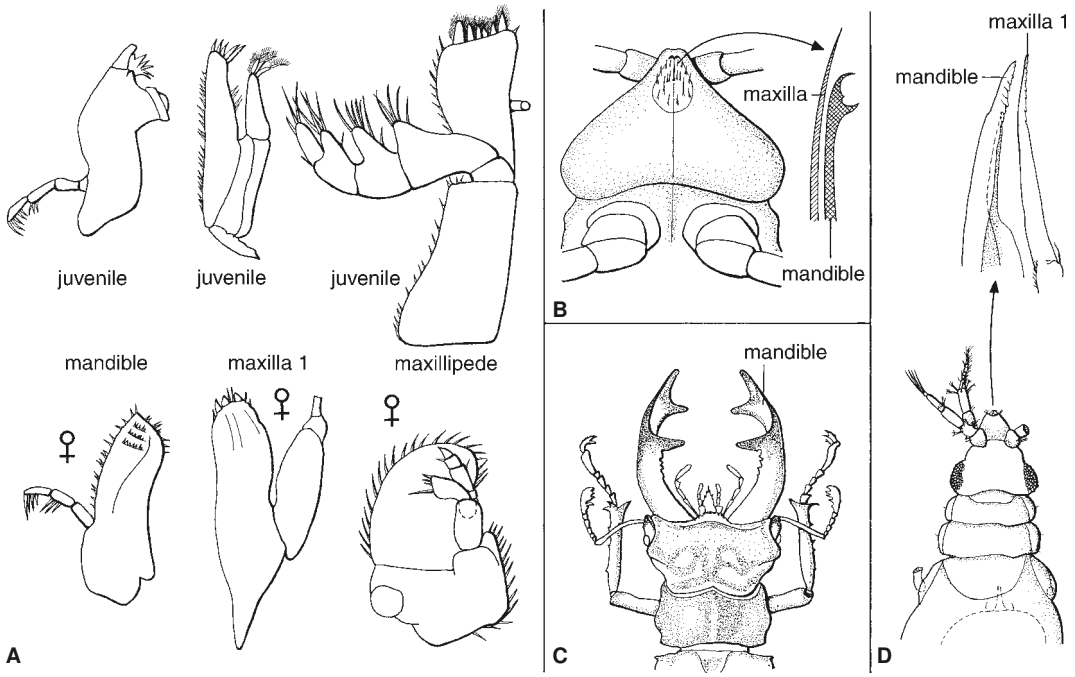


Fig. 101. The hypothesis of homology of the mandible of Crustacea and Tracheata is independent of the function of the mandible. **A.** The chewing mouthparts of *Dynamenella curalii* (Crustacea: Isopoda) are only normal and functional in immature animals, in mature females they are atrophied and without function. **B.** Piercing mouthparts of *Neanura muscorum* (Collembola). **C.** The mandibles of male *Lucanus cervus* (Coleoptera) are relatively harmless weapons of attack, used to drive out other males. **D.** Piercing mouthparts of Gnathiidae (Crustacea: Isopoda) are only present in juveniles *(Caecognathia calva).*

in *Artemia* first in the leg bud, later in primordia of epipods (the gills), *ap* additionally in the leg musculature.

4) The expression of the same genes proves the homology of epipods (gills) of crustaceans and wings of insects.
5) It follows that all primarily wingless Tracheata reduced epipods convergently. In pterygote insects, however, they evolved to wings.

In this argumentation there are several inaccuracies. (a) The embryonic anlagen of insect legs develop in the pleural region; and at the border between pleura and tergite the wings also develop. The spatial proximity does not imply a homology. (b) The genes studied are not unique to crustaceans and insects, and gene expression is not limited to the epipod in crustaceans, thus the genes and their expression cannot be specific for epipods. (c) Homology due to structural identity has only been shown for the genes themselves, not for the factors triggering their transcription and for the following cascades. Even when the transcription factors were homologous, conclusion 4) would not be acceptable, because a specific protein can be needed in different non-homologous parts of the body. (d) Finally, conclusion 5) is not plausible, because the epipods (= gills of crustaceans) would have no function on land and therefore it would be highly improbable that they were retained during evolution of insect ancestors up to the first Pterygota with functional wings. The example of other terrestrial animals with aquatic ancestors (Pulmonata, Amphibia, Chelicerata) shows that with the adaptation to life on land gills are either reduced or internalized. Furthermore, nowhere in extant wingless Tracheata exist remnants or embryonic anlagen of epipods which could support the theory of homology of wings and epipods.

We see that expression of the same genes is not necessarily an indication of the homology of the corresponding body regions. Evidence for the expression of homologous genes in non-homologous organs exists also for homeotic genes. Homologues of the *distalless* transcription factor, for example, are expressed in developing insect legs and appendages of other Articulata, in fin and leg primordia of Gnathostomata, but also in legless Tunicata and in podia of echinoderms. The

hedgehog gene is not only expressed in leg buds of arthropods and vertebrates but also in anlagen of gill arches of chickens (Shubin et al. 1997). The *fringe* gene is expressed along the edge of the wing of insects and on the wing of birds (Gaunt 1997). Nevertheless, wings of birds are not homologous to those of the Pterygota. Due to the same reasons the assumption that metamerism of vertebrates is homologous to that of insects is unfounded (hypothesis of De Robertis 1997).

The **criterion of congruence** has already been introduced (see ch. 5.1.1). The distribution of characters mapped after a cladistic construction of a most parsimonious dendrogram onto stem lineages serves as evidence of homology. This criterion is only valuable when the characters have been previously weighted according to their probability of homology and when the algorithm used to construct the tree considers these weights. Criticism of this criterion is summarized in ch. 6.1.10.

When the evolution of a character or a gene is observed directly, as it can occur with breeded animals or with cultures of bacteria, it is possible to use the **criterion of common descent**. In most cases, however, data gained through direct observation are not available for systematics. Therefore, common descent is only a theoretical explanation for observed patterns.

Sometimes it is claimed that it is necessary to know the **function** of organs in order to be able to support a hypothesis of homology. A mandible, for example, is a special instrument originally adapted for chewing in insects, myriapods or crustaceans and for functional reasons it is placed laterally to the mouth opening and covered by the labrum frontally. The similar functional integration is said to be an evidence of homology. However, we have to realize that **function is irrelevant**, because also the piercing-sucking mandible of some insects or the vestigial, "useless" mandibles of some adult crustaceans (Fig. 98A) are homologous to the original chewing mandible in the sense that it is a character inherited from the ancestral Mandibulata. Should a crab be discovered which possesses a mandible transformed into a paddle used for swimming, this structure would nevertheless be homologous to a "normal" mandible.

The **rejection of hypotheses of homology** results from the preceding statements. In many cases we simply have to admit that we have no evidence in favour of a hypothesis of homology. But there exist also suggestions for a special justification of a rejection. The **criterion of conjunction** (Patterson 1982) states that two similar structures occurring simultaneously in the same organism cannot be homologous. Patterson constructs the following example: if there existed angels and were the wings of these angels homologous to those of birds, then the hypothesis that wings of birds are homologous to the front legs of mammals would have to be rejected, because angels have wings *as well as* arms. This argumentation is not correct: there are enough examples for duplications of organs in animals and plants. **Homonomy** is the fact that homologous organs occur multiple times in a single organism (metameric internal organs of Articulata, vertebrae of vertebrates, the large number of legs in Notostraca, Pantopoda, but also hairs, erythrocytes, etc.). The criterion of conjunction is not useful. For the rejection of hypotheses of homology there exist only the following arguments:

- the characters do not show correspondence in detail and are only superficially similar, while details are present also in other taxa (eye of a squid compared with eye of a fish and eye of *Nautilus*).
- there are characters of higher probability of homology, whose distribution among taxa is incompatible with the distribution of the character in question (e.g., shared reduction of pigmentation in cave spiders and cave crickets in comparison to the occurrence of spinnerets in cave spiders and pigmented spiders: example of Fig. 136). A prerequisite for this argument is that the complex character does not spread by horizontal gene transfer.

To choose from potentially homologous characters, systematists offer very different recommendations (Fig. 102).

It follows inevitably from these arguments (Fig. 102) under which circumstances the support of a hypothesis of homology must be considered to be weak, for example when characters evolve so fast that ancestral characters are not conserved, or convergences are expected to occur frequently, or when characters are not complex enough, and

criterion	argument	evaluated fact
Conserved or slowly evolving characters are better (Sober 1986)	Fast evolving characters quickly get "noisy", with the consequence that autapomorphies and analogies evolve more often and synapomorphies erode	evolutionary process
Complex characters are better (e.g., Hennig 1950, Remane 1961)	The probability of homology is higher	complexity of patterns
Characters without a function are better, adaptive characters are less suitable (Darwin 1859)	Convergences are less probable	evolutionary process
Characters that are functionally independent of each other are better (e.g., Mayr & Ashlock 1991)	In functionally dependent characters, substitutions of one detail may depend on previous ones in another detail, the weight of similar characters may be overestimated	evolutionary process
Characters occurring with others in the same species are more important	The probability of homology is higher (criterion of compatibility)	complexity of patterns
Reductions are weak characters	Convergences can occur frequently. Or: characters referring to reductions are not complex	evolutionary process or complexity of patterns
Polymorphic characters independent of sex or caste are not suitable (Darwin 1859, Simpson 1961, Wiens 1995 etc.)	When polymorphic characters are not present in a ground pattern of a taxon, they are not characters of the taxon, but instead of individual organisms; a homologization comparing different taxa is not possible	complexity of patterns

Fig. 102. Table with criteria for the selection of "good" characters.

whenever the relative probability of homology cannot be estimated due to lack of data.

Characters of different developmental stages

Is it true that morphological characters can only be homologized when they are compared at the same developmental stage (for example, in the third naupliar stage or at the moment after birth)? To take stages of ontogeny into account is not important whenever the complexity of a character allows homologization. Heterochrony (Fig. 114; a change in developmental timing) may cause the development of an organ in a different stage than usual. Remember that the occurrence of a complex structure indicates the presence of specific genes, which may be activated at different times during ontogeny. An organ is also homologous when the corresponding genes are transcribed in different phases of ontogeny. An anlage of mammalian teeth is an indication of the presence of "tooth genes" whether it occurs in an embryo within a whale's womb or in a human baby that is already some months old. Therefore Nelson's rule for recapitulated characters is relevant to identify plesiomorphic homologies (ch. 5.3.5). The real problem lies in the estimation of the certainty for the detection of the presence of homologous gene variants. Difficulties arise with characters of low complexity and therefore uncertain homology. When the number of bristles on the appendages of crustaceans or spiders is used as a character to differentiate species, then size, sex and age of the animals must be considered, because with increasing size the number of bristles usually increases too. The number of bristles is a "weak" character.

Attention: It has to be pointed out precisely which details are considered to be homologous. The fact that anlagen of gill slits occur in embryos of

mammals does not mean that gills of adult fishes are homologous to these anlagen. The latter do not show the fine structure of gills. Only the details shared by both structures can be homologized.

A totally different question is whether the point in time of occurrence of a character during ontogeny can be a homology. For example, the eruption of milk teeth occurs later in anthropoid apes than in other Catarrhini. This detail belongs to an extensive pattern in space and time, the retarded postnatal growth of anthropoid apes. This pattern serves as evidence for the presence of homologous genetic characters, assuming that the point in time of a morphogenetic process is regulated by homologous genes. In this case a developmental pattern is a frame homology and it is necessary to present arguments supporting homology of changes found within this frame, which is not a simple problem. The suppression of larval stages, for example, occurs frequently in lecithotrophic marine invertebrates and is a character with no specific complexity.

May characters be ignored for a phylogenetic analysis?

At the end of this section this question is raised, because a criticism by cladists working phenetically is that some phylogeneticists choose information subjectively and reject other data without objective reasons. The answer to this question is clearly "**yes**": one must not and shall not take into account those characters which have a very low probability of homology, because such characters can be numerically dominant and introduce only "noise" and wrong signals into the analysis. This is especially important when only a limited number of characters are used, as in comparative morphology. But the answer is also "**no**": ignoring known characters of high probability of homology is an act of ignorance or of manipulation of results.

5.2.2 Homologization of molecular characters

For the homologization of molecules basically the same rules apply as for morphological characters although sequences allow a more exact quantification of identical details and more mathematical procedures can be used for pattern analyses. Homologies of nucleic acids can be found on very different levels. The following possibilities are used in practice:

- determination of the homology of a sequence position with the help of alignment methods (ch. 5.2.2.1),
- phenomenological determination of the homology of a sequence region (ch. 5.2.2.2),
- phenomenological determination of the homology of a gene (ch. 5.2.2.3),
- phenomenological determination of the homology of a gene arrangement (ch. 5.2.2.3),
- phenomenological determination of the homology of duplicated sequences (ch. 5.2.23),
- evaluation of distance data of DNA-DNA-hybridizations (ch. 5.2.2.8).

Other molecular data allow the assessment of the similarity of metabolic products, of allozymes, of amplification or restriction fragments, of amino acid sequences. On principle it has to be taken into account that DNA-sequences have the advantage to be present in each cell, whereas proteins and other metabolic products are often only produced in specific organs or during specific physiological states. Absence of a metabolic product cannot be equated uncritically with the lack of coding genes belonging to the corresponding metabolic pathway.

5.2.2.1 Sequence alignment

When DNA-sequences are copied without modification, the copies are homologous in every detail, including the order of individual nucleotides. If one nucleotide has been exchanged by another one, then the novelty occurs at the position of the predecessor. The new nucleotide is *not* homologous to the old one, but the **position** is the same. For this type of comparison, the single sequence position is equivalent to the frame homology as defined in comparative morphology (ch. 4.2.2 and 4.3.1). The frame homology must be established prior to the reconstruction of the phylogenetic tree, because most methods of tree construction do not imply a test of frame homologies (but see also optimization alignment, ch. 14.12).

The correct alignment is useful for several purposes. It will show which positions have been

constant during evolution and which vary. Comparing these patterns with the three-dimensional structure of a protein it is possible to identify the regions that are of functional importance and therefore conserved, such as catalytic areas. The alignment will also show if there exists length variability, if nucleotides were inserted or deleted in some of the sequences. However, inverted or translocated sequence areas are not automatically detected with alignment programs; for this purpose it is necessary to search for sequence motives.

Longer sequence sections or genes are with high probability homologous, whenever they show the same order of nucleotides in most parts of the sequence (see analogy of identical words, Fig. 87). As in phylogeny inference nucleotides are compared position by position, the homologization of such sequences requires that the most probable **positional homology** has been determined correctly. This is done with the help of **alignment methods**, which arrange sequences in rows of a matrix in such a way that each presumably homologous position is found in the same column of the matrix.

Longer homologous sequences or genes may also contain non-homologous regions, which originate for example by loss and/or insertion of nucleotides or of longer sequence fragments. Sequences can be of different length due to insertions or deletions. Therefore gaps have to be introduced in shorter sequences to allow that all homologous positions are written in columns. After alignment all sequences have the same length (Fig. 103). Gaps represent those positions in which other sequences show insertions or where nucleotides were lost. These gaps cause uncertainties when selecting the best alignment: often there exist more than one alternative to choose from.

The importance of the alignment procedure must not be misjudged, it is one of the decisive steps in the analysis of phylogeny with sequences. It is absolutely equivalent to the *a priori* determination of homologies in comparative morphology. Modifications of alignments of the same dataset can have a larger influence on the result of phylogenetic analyses than alternative methods of tree reconstruction (Morrison & Ellis 1997). When sequences contain many gaps, simple variations

	alignment 1	alignment 2
species A	G C G T A A T	G C **G** T A A T
species B	G **G** - T A A T	G - **G** T A A T
species C	G **G** T T A A T	G G T T A A T

Fig. 103. Example for two alternative alignments of the same sequences. Nucleotides of ambiguous homology shown in bold letters.

of the alignment (Fig. 103) can produce very different dendrograms.

There are different methods to optimize the alignment, research in this field is not yet concluded. The fundamental problem of all alignment methods is that an efficient algorithm does not necessarily reconstruct the historical processes of insertion, deletion and substitution (Fig. 104).

Many popular algorithms used for the optimization of alignments proceed according to the principle that during the adjustment of the length of different sequences, certain steps are punished with points and the alignment with the lowest number of points is chosen. A step may be: the insertion of an alignment gap, or the retention of a variable position. Steps can be weighted, for example, by defining the first opening of a gap as more "expensive" than the elongation of the gap, or by giving the presence of a transition lower weight than transversions when comparing two sequences. Algorithms are explained by Needleman & Wunsch 1970, Waterman 1984, Davison 1985, Lipman & Pearson 1985 and others (texts dealing with alignment algorithms can also be found in the internet: ch. 15). Furthermore, the number of constant positions (positions with the same nucleotide in each sequence) and the number of different characters per position can be counted in order to characterize alignments.

In alignments of DNA sequences coding for proteins, the reading frame of the coded amino acids has to be taken into account in order to determine the probability of homology of nucleotides (see Altschul 1991, Claverie 1993). Coding DNA sequences are usually aligned pairwise like non-coding ones. For the evaluation of different alignment variants, the number of identical nucleotide pairs, the chemical properties of the coded amino acids, and the number of alignment gaps can be considered. Another possibility is to count the codon changes that result in the coding of a new amino acid. More often empirically ascertained mutation frequencies are used in form of weight matrices (e.g., Dayhoff-Matrix, Gonnet matrix, see ch. 5.2.2.10). These have the effect that amino acid changes that are rare are more expensive. When available, the fit to known secondary and tertiary structures of the folded molecules can be considered.

Starting from algorithms which find the most parsimonious alignment of two sequences, methods have been developed which align along a dendrogram: either the topology is given (e.g., on the basis of morphological characters and a previous phylogenetic analysis) or a first topology is calculated from the raw data, using pairwise alignments and a distance tree estimated from pairwise distances (e.g., with the computer program CLUSTAL, Higgins & Sharp 1988, Thompson et al. 1994). The tree is used to guide the multiple alignment. Attention: the order in which the sequences are read often has an influence on the result of many alignment algorithms! Using CLUSTAL, the result depends on two important parameters that have to be chosen: the **gap opening penalty**, and the **gap extension penalty**. If both parameters are low, more gaps are introduced in the alignment. After changing these parameters one can examine if the alignment improves: the number of conserved positions should be as high as possible.

Dendrograms for alignment procedures can also be calculated with other methods (Maximum Parsimony (MALIGN: Wheeler & Gladstein 1994; POY: Wheeler 1996), Maximum Likelihood). The first topology determines in which order the sequences are aligned further, optimizing the alignment for the most similar sequences (sister taxa in the chosen topology) first. After the first step the sequences that have been aligned first are represented by a consensus sequence, and the next similar sequence is included in the alignment and adjusted to the consensus sequence. In principle, this method can be performed iteratively since after the conclusion of the first alignment a dendrogram can be calculated and this is used for a descending analysis, starting again with the alignment of the most similar terminal sequences. The procedure is repeated until no further changes occur. An advantage is that parsimony-trees and alignments are constructed simultaneously. Disadvantages are that computation time is considerably increased and the effect of unalignable (ambiguous) areas is obscured. Optimization alignment is explained in more detail in ch. 14.12.

A promising idea is not to use a dendrogram as the basis for the selection of the first sequences to be aligned, but to start with a pre-aligned dataset with an approximation method (e.g., CLUSTAL, Higgins & Sharp 1988) to search for the splits with best support, and then to modify the alignment optimizing the phylogenetic signal in favour of these splits. Then further splits can be searched within the optimized splits and in those positions that have not been considered so far.

A totally different approach was proposed by Morgenstern et al. (1996). The individual nucleotides are not shifted and alignment gaps are not weighted, but complete sequence fragments that are highly congruent are searched for. These fragments can be aligned, whereby a weight that depends on the similarity of the fragments of two sequences decides which fragments are aligned first. In further steps the less congruent regions are also aligned, while the result of the alignment of the fragments "of first choice" is maintained unchanged. Gaps do not have to be inserted explicitly, they are only the regions remaining between aligned fragments.

Experience shows that computer programs produce good alignments mainly for sequence regions which show little variability. Hypervariable regions are also aligned, however, the result is usually worthless. Moderately variable regions can sometimes be re-aligned "by hand" to optimize the number of invariable and possibly homologous positions. Attention: alignment algorithms often make obvious mistakes that are easily noted by eye.

Recommendations for practical work:

- an alignment calculated by computer should always be controlled visually (as long as no better algorithms exist) in order to test whether very variable sequence regions which were not aligned optimally or which simply cannot be aligned occur.
- Very variable regions, for which many equivalent alignment alternatives exist, have to be eliminated from the dataset, or have to be ignored when calculating dendrograms, because the probability of homology is very low.
- The uncontrolled elimination of all positions showing alignment gaps can lead to a heavy loss of informative positions. At the moment one should rely either on methods that compare variability of positions and tree topologies or on the variability visible by eye. It is wise to select for tree construction areas for which no or only few alignment alternatives are recognizable.

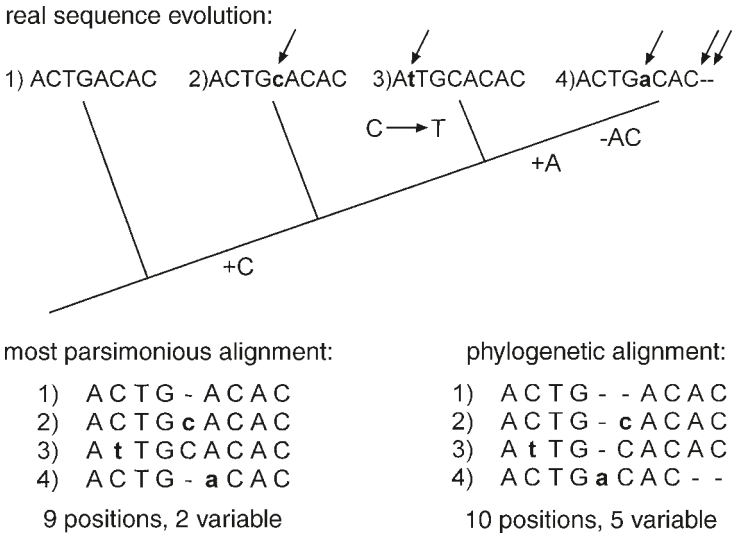


Fig. 104. Assuming that the evolution of a sequence proceeded as in this example, the most parsimonious alignment may not necessarily be the historically correct one. Novelties which did not occur in sequence 1 are shown with small letters.

- When several optimal alignments are found, it should be tested whether different dendrograms can be reconstructed with them.
- When very different sequences are aligned, better results are obtained aligning groups separately. To do so, small groups of most similar sequences or smaller taxa of established monophyly are processed individually, and only afterwards these groups are aligned maintaining positional homology within groups ("profile alignment"). In this way a better approximation to the phylogenetically correct alignment is obtained. The most parsimonious alignment is not necessarily the correct one (see Fig. 104).
- If a well supported phylogeny is known, algorithms should be chosen which allow alignment along a given topology.
- When monophyletic groups are not known *a priori* for large datasets, algorithms would be useful which do not align along a topology but instead optimize the alignment for splits with the highest signal. (Such methods still have to be developed!).
- If the secondary structure of rRNA sequences is known, helical regions should be aligned separately. There exists software that allows the search for matching positions in helical regions (DCSE: De Rijk & De Wachter 1993, unfortunately not very user-friendly).

The homologization of sequence positions is often difficult because the individual character (the position) does not show any complexity. Certainty for the correct homologization can only be gained with the search for identical patterns of nucleotides in a sequence region (see next chapter). It is indeed the pattern of neighbouring nucleotides that determines the alignment of single positions. In the example of Fig. 103, the pattern "TAAT" is homologized for three sequences. The longer such a conserved sequence section is, the smaller is the probability that the identity is a product of chance.

The output of alignment programs can have different formats, depending on the software used. There are two main variants:

The **sequential format** is composed of blocks, each consisting of a complete sequence:

```
Species_1   TAATTAAAGG GCCGTGGTAT A-CTGACCAT GCGAAGGTAG CATAATCATT
            AGCCTTTTGA TTTGAGGCTG GAATGAATGG TTTGACGAGA GATGGTCTGT
            CTC--TTCGA T-TAAATTGA AGTTAATCTT TAAGTGAAAA AGCTTAAATG
            TACTTGGAGG GCGATAAGAC CCTATAGATC TTTACATTTA AT-TCTTTTG
            TCTTGCGGTA G-GTAATTAG ACAGAGTA-- -AAACA---- --ATGTTCGG
            TTGGGGCGAC GGTAAGA--A CAGAATAAAC -ACTTACAAC ATAAACACAT
            CAATAAATGA CCA------- TTGATCCT-T AGATGAAT-- AAAGACCAAG
            TTACCTTAGG GATAACAGCG TAATTCTTTT TTGAGAGTTC AAATCGACAA
            AAGAGTTTGC GAGCCTCGAT

Species_2   GTG..G.... ....C..... TTT.....G. .......... ..........
            ..T.....A. ..GA.....C .......... ..G.....A. A.CACA....
            T..AG--AC. AAG..T.... .T....CT.. .......... G.........
            A.T.AAA.A. A......... ......A... ....----. T.A.G.....
            .A..TTTAA. .TTGTTG.GT .TTA.AA.GA A.T.T.AAGT AG.....GT.
            .......A.. .A...T.TA. T.AG..T.G. TGT.GGT..A ..T..ACA.A
            TGTGT-T..T TAGGAGTAGT .A......T. TT.GAG..TT .....TT...
            ....T..... .......... .T...-.... C......... TT.....AG.
            ..A..A.... ..-.......

Species_3   AT...G.... ....C..... TT......G. .......... ..........
            ..T.....A. ..G......C .T.....A.. .CGA.....T ...A.A....
            ...AGCAAA. AAA....... .C....CT.. A......... G....T...A
            .TTCA.A.A. A......... ......A... ..C..----. CCAC.....A
            .AC.AA.CC. ATC.GT..GT .T-A..G.GA AGT.T.AAAA ACG....TT.
            .......A.. AAAG.T.TA. TTA...-... TGT.--TT.A T.TT.A.ACA
            A..AT-T.-T TG.AAATAAA ........C. CT.AAG..TA .....T....
            ....T..... .......... .C...-...C .......... CC.....A.G
            ..A....... ..-.......

Species_4   A........A ....CA.... TT-....TG. .......... ......A.A.
            T.T.....A. A.AA.....T .......A.. ..G...A.AG T..CA.....
            T..--..AA. .A.TT..... .T..G.CT.. .......... G........A
            A.TCAAA.A. A......... ......A... ......A.A. ..A.A....A
            .A..TTA..T -TA...G.GA .TGATA..TA A...T.TAGG TT-....GC.
            C.......G. .TAG.T.TA- -TA....... TGTC..T.GT T....T...A
            T...C.T..C TT.ATATAAA .........-. .A..AG..TA .....TT...
            A...T..... .......... .T...-.... .......... ..........
            .......... .......... .
```

In this example, the dots represent the same character state as in the first sequence.

The **interleaved format** shows in each line the same positions for different species:

```
Species_1   TAATTAAAGG GCCGTGGTAT A-CTGACCAT GCGAAGGTAG CATAATCATT
Species_2   GTG..G.... ....C..... TTT.....G. .......... ..........
Species_3   AT...G.... ....C..... TT......G. .......... ..........
Species_4   A........A ....CA.... TT-....TG. .......... ......A.A.

Species_1   TTTGACGAGA GATGGTCTGT CTC--TTCGA T-TAAATTGA AGTTAATCTT
Species_2   ..G.....A. A.CACA.... T..AG--AC. AAG..T.... .T....CT..
Species_3   .CGA.....T ...A.A.... ...AGCAAA. AAA....... .C....CT..
Species_4   ..G...A.AG T..CA..... T..--..AA. .A.TT..... .T..G.CT..
```

```
Species_1    GCGATAAGAC CCTATAGATC TTTACATTTA AT-TCTTTTG TCTTGCGGTA
Species_2    A......... ......A... ..........- T.A.G..... .A..TTTAA.
Species_3    A......... ......A... ..C..-----. CCAC.....A .AC.AA.CC.
Species_4    A......... ......A... ......A.A. ..A.A....A .A..TTA..T

Species_1    --ATGTTCGG TTGGGGCGAC GGTAAGA--A CAGAATAAAC -ACTTACAAC
Species_2    AG.....GT. .......A.. .A...T.TA. T.AG..T.G. TGT.GGT..A
Species_3    ACG....TT. .......A.. AAAG.T.TA. TTA...-... TGT.--TT.A
Species_4    TT-....GC. C.......G. .TAG.T.TA- -TA....... TGTC..T.GT
```

5.2.2.2 Determination of the homology of nucleotides and of sequence sections

It has often been stated by several authors that the sequence position is the character, the nucleotide is the character state. One could as well say that the position is a frame homology also including the neighbouring regions, the nucleotide is the detail homology within this frame. The identity of a single nucleotide in two or more sequences in an optimal alignment is not a good indication for homology. The example (Fig. 105) shows in the same alignment incompatible identities (homoplasies, inconsistencies), which support two different partitions (**splits**), of which at best only one can be the trace of a real speciation.

Differences in probability of homology can be considered by **weighting,** as with morphological characters (ch. 5.1.2, 6.1.3). Weighting of nucleotides is possible in different ways (see also Brower & DeSalle 1994):

- Phenomenological *a priori* weighting: evaluation of single nucleotides as parts of a complex pattern, for example, with a phenomenological "spectral analysis" (ch. 6.5), in which all partitions represented in a dataset are filtered out separately. The positions supporting the strongest signals of a spectrum can get the highest weight. Furthermore, the variability of individual positions as visible in the alignment can be used for weighting.
- Phenomenological weighting with the simpler, not so strictly justified combinatorial weighting (appendix 14.2.2).
- Model-dependent *a priori* weighting (s. ch. 6.3.1, ch. 7, Fig. 155): weighting of the probability that certain substitutions occur, for example, transversions or specific substitutions like G⇒C (Fitch & Ye 1991), or weighting of compensating substitutions in helical RNA-sections, weighting with model-dependent Hendy-Penny spectra (ch. 8.5, 14.7). Also the evaluation of selection pressure (e.g., putting low weight on the third codon position: ch. 6.1.2.5, ch. 6.3.1) is a form of model-dependent weighting.
- *A posteriori* weighting: evaluation of the character distribution in an inferred dendrogram, for example with successive weighting in the MP-method (ch. 6.1.4), which is analogous to the circular (!) weighting of morphological characters.

Indirect weighting is done with maximum likelihood methods, where probabilities are assigned to character states according to their distribution in a tree.

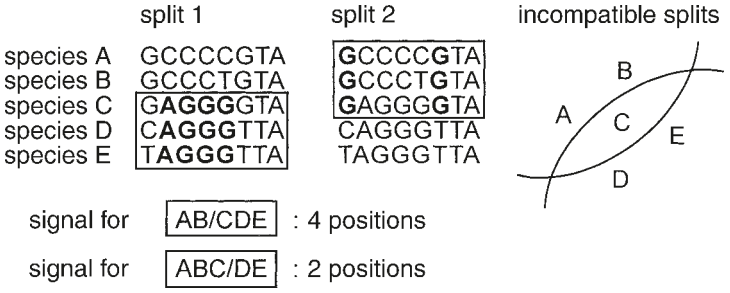


Fig. 105. Incompatibility of two signals.

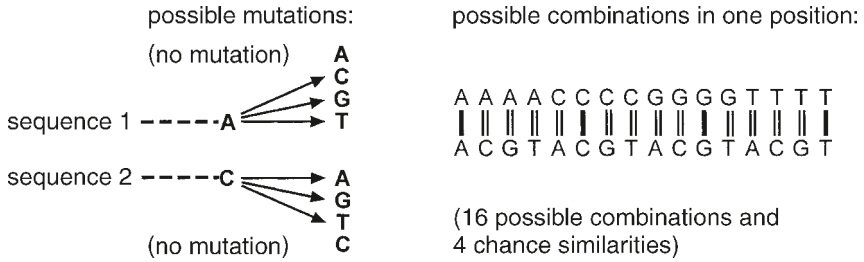


Fig. 106. What is the probability for the occurrence of chance similarities in a position of two unrelated DNA-sequences or for the occurrence of similarities after random mutations?

Model-dependent weighting is performed without statements on the individual case ("position 1131 mutates from A to C"), as explained in the chapters on distance and maximum likelihood methods (ch. 8). Using these methods an assessment of homologies according to the probability of cognition is not relevant, but one needs statements on the probability that certain evolutionary processes occur. Rules that govern the evolution of sequences are not well known, evolutionary processes might often be irregular, wherefore assumptions on the probability of a character transformation which allow good reconstructions for one taxon might be misleading in other taxa or for other genes.

Within the framework of model-dependent methods weighting is also done on the basis of base frequencies (nucleotide frequencies) in terminal sequences (the relation A : G : C : T is not necessarily 1 : 1 : 1 : 1) and considering the frequency with which specific substitutions occur in a dendrogram (the number of substitutions A ⇒ C and C ⇒ A may be different) (ch. 8.1, Fig. 44). This type of weighting can also be used in MP-analysis (Fitch & Ye 1991) (see ch. 6.1.3).

A posteriori weighting of nucleotide patterns is equivalent to the one for morphological characters in the framework of phenetic cladistic methods (ch. 6.1.3, 6.1.4). This circular method is not admissible.

Sequence sections

Considering a single sequence position of two aligned DNA-sequences (Fig. 106), the probability that the same nucleotide occurs by chance after mutations is about ¼ (assuming base frequencies are equal). For longer sequences the probability that all nucleotides are similar in all positions is reduced and is with n positions 4^{-n} (when all nucleotides and all types of substitutions have approximately the same frequency). Identity by chance of a sequence of 5 positions would occur with a probability of about 10^{-3}, it would be much rarer than in the case of only one position.

For this reason it is to be expected that **signature sequences** of 10 or more nucleotides can be apomorphies of high probability of homology. So far only few such sequences have been searched for directly. They can be found in coding as well as other DNA regions. For example, the Cirripedia (barnacles and related crustaceans) have a characteristic insertion with the sequence CTGGGCTCCC in their 18S r DNA (Spears et al. 1994, Fig. 98). Microsporidia have in their EF-1 peptide a sequence of 10 amino acids, which also occurs in fungi and in Metazoa. This is evidence against the placement of the Microsporidia at the base of the Eukaryotes (Philippe & Laurent 1998). Signature sequences are used frequently for the **PCR technique**: homologous genes are amplified with the help of primer molecules, which are so long that a match with non-homologous sequences is very improbable. For example, the sequence 5'-CCTACCTGGTTGATCCTGCCAGT-3' can be used as primer for the identification of the beginning of the 18S rRNA-gene.

Interesting markers are the SINEs (short interspersed nuclear elements). Their function is not known, they constitute 5-10 % of the mammalian genome. Although these sequences should evolve neutrally were they really free to vary, in reality they are highly conserved and can be retained in the genome for at least 100 million years. The sequence MIR (GCCTCAGTTTCCTCATC)

allows hybridizations or PCR-amplifications in mammals and is also found in birds; the sequence ARE-A2 (ACTGAATCACTTTGCTGTACAG) only occurs in ungulates and whales, BOV-A2 in ruminants (Tragulina + Pecora, not in Tylopoda and Suiformes) (Buntjer et al. 1997).

5.2.2.3 Homology of genes, gene arrangements, sequence duplications

Whereas the analysis of nucleotide sequences yields a large number of characters with low weight, each character being one of four nucleotides, specific gene arrangements are more complex characters. The complexity is based on the larger number of "letters" of which the patterns are composed. In this case the letters are genes which can occupy different positions on a chromosome.

Loss mutations of larger sequence sections are as unspecific as for morphological characters, they are not complex characters and can occur convergently (e.g., convergent loss of repetitive cpDNA sequences in conifers and leguminoses (Lavin et al. 1990) or of cpDNA introns in several flowering plants (Downie et al. 1991)). They should not receive a high weight.

Gene rearrangements can be valuable apomorphies when complex changes can be tracked down. The transposition of a single gene of the mtDNA is an event for which analogy cannot be ruled out, the combination of several characters allows a safer statement of homology. In breakpoint analyses, the homology of gene order is determined *a posteriori*, i.e. by mapping of characters on the optimal tree.

Breakpoint analyses require the definition of a distance between genomes based on the number of break points and is applicable when the number of genes does not change. A breakpoint is a difference between two genomes caused by the presence of two neighbouring genes in one genome and the absence of this pairing in the other genome (Blanchette et al. 1997, Wang 2002). Existing algorithms are computationally very intensive.

Examples for changes of gene arrangements: complex modifications have been shown for lizards. The monophyly of the Acrodonta in comparison to the sister group Iguanidae is supported by the reduction of the replication origin O_l for the light strand of the mitochondrial DNA. Furthermore, the tRNAs for glutamine and isoleucine have swapped places and the cystein-tRNAs have lost the D-helix (Macey et al. 1997). The coupled occurrence of these characters increases the probability of homology. – Birds as well as snakes and lizards have in the mitochondrial genome tRNA-genes in the order $tRNA^{His}$ – $tRNA^{Ser}$ – $tRNA^{Leu}$, whereas crocodiles show the sequence $tRNA^{Ser}$ – $tRNA^{His}$ – $tRNA^{Leu}$. This shift is an apomorphy of crocodiles (Kumazawa & Nishida 1995). – Starfish have the mitochondrial gene order 16S-ND2-ND1, while in sea urchins the sequence is reversed (Asakawa et al. 1991).

How often changes of gene arrangements occur and whether specific translocations are favoured by cellular mechanisms is not known. In the mitochondrial DNA of humans anomalies were frequently detected, like the duplication of some sections, which seemingly do not have any consequences for the affected persons. This indicates that such mutations are not rare (Tengan & Moraes 1998). If such mutations are frequent a homology for single identical mutations is less probable.

Transpositions of sequences from organelles into the nucleus are relatively rare events. When sequences which originate from organelles are found in the nucleus, the homologization of a transposition is possible on the basis of the identity of nucleotide patterns of the insert and the site (neighbouring sequences) where it is found. Such transpositions are known for nucleic acids of chloroplasts and mitochondria (Blanchard & Schmidt 1995, Zischler et al. 1998) and are in important evolutionary mechanism (Martin et al. 1998). For example, in the cell nuclei of gibbons, orangutans, gorillas, chimpanzees and in humans occurs an insertion of about 360 nucleotides which is a copy of part of the mitochondrial DNA (D-loop) (in humans on chromosome 9). With this insertion an apomorphy of the Hominoidea has been discovered which is absent in other primates: the event must have occurred in the stem lineage of the Hominoidea (Zischler et al. 1998).

Duplications can affect sequence regions, complete genes, chromosomes, or whole genomes. Gene duplication followed by mutation and se-

lection is an important mechanisms in the evolution of biological diversity. Gene duplications can produce redundant genes which all have the same function and are needed by the cell in high copy number (repetitive genes). For example, in Metazoa there are up to thousands of tandem repeats (depending on the taxon) of rDNA genes in each cell, which may occasionally show individual mutations, but are generally "homogenized" after few generations through losses and duplications ("concerted evolution": s. Elder & Turner 1995). These processes proceed so rapidly that in the individual organisms different copies of different duplication events are usually not detectable. The genes are primarily serially homologous to each other. When the genes evolve further independently after the duplication, orthologous and paralogous genes can be distinguished (see Fig. 7, ch. 4.2.2, 4.2.4): after gene A is duplicated, the daughter genes A1 and A2 continue to evolve and may gain different functions. Both gene copies can occur in different species: gene A1 in individual X is orthologous to gene A1 in individual Y, but paralogous to each copy of gene A2. After duplication genes may specialize for different functions.

Useful characters result from rare duplications. The duplication event itself as well as the occurrence of novelties in the further course of evolution of gene copies provide phylogenetic information. Repeated duplication events can create "gene families" composed of independently evolving paralogous gene copies which can be characterized individually on the basis of unique mutations. It is important to distinguish paralogous and orthologous genes in phylogenetic analyses: gene sequences isolated from different individuals may not be orthologous (homologous in the stricter sense) but may belong to two different lines of paralogous genes. In a phylogenetic tree these lines do not necessarily merge where the species split, but at an earlier point in time, the moment of gene duplication (s. Fig. 6). In this case the gene tree may differ from the species tree. Orthologous and paralogous genes are recognized on the basis of unequal sequence differences: paralogous genes are less similar to each other than orthologous genes. This distinction requires the comparison of several sequences; ideally, to understand gene evolution all paralogous variants should be known.

Examples: Hox proteins with homeodomains (a highly conserved 60 amino acid DNA-binding domain) are transcription factors having important functions in the regulation of embryogenesis. Two groups of proteins which evolve separately originated through gene duplication in eukaryote ancestors of higher plants, fungi and Metazoa (Bharathan et al. 1997). The mammalian Hox gene complex is a group of genes (39 in mice) which are located on 4 linkage groups. Each linkage group contains copies of the same set of maximally 13 genes. These genes regulate during embryogenesis patterning along the rostrocaudal axis and each type of gene has regulatory functions in a specific region of the body (e.g., *Hox1* genes are expressed in the most anterior part of the body; Carroll et al. 2001). – In mammals, paralogous hemoglobin genes that are specific for embryonic, fetal and adult life stages differentiated after gene duplication. – Examples for further gene families: tubulin genes of Metazoa, actin genes, hemoglobin genes of vertebrates, serpins, insulin-like genes, kinesins, enolases, chaperonins, ABC transporters, aquaporins, etc.

5.2.2.4 Homology of restriction fragments

Restriction fragments are obtained through treatment of nuclear or mitochondrial DNA with restriction endonucleases. These cut DNA molecules at sites with specific nucleotide sequences (recognition sites), with the effect that the treatment of the same DNA region with the same enzymes always yields the same fragments (protocols can be found, e.g., in Dowling et al. 1990). Even when long sequences are compared, specific information on mutations is only obtained in those small sequence regions which correspond to the recognition sites of the enzymes (usually less than 1 % of sequence length), because a mutated recognition site will not be cut. Furthermore, length variation of homologous fragments will be visible after gel electrophoresis. As many homologous cleavage sites may be retained without modification in the course of evolution, only the few variable ones are valuable as phylogenetic markers. It is possible to homologize

a) the cleavage sites and
b) the fragments.

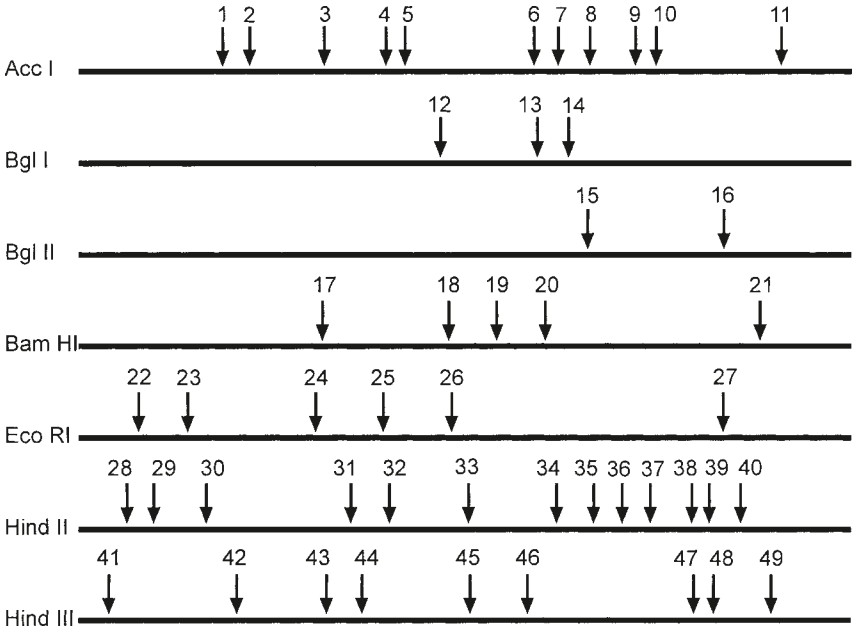


Fig. 107. Map of cleavage sites found in European hares (Pérez-Suárez et al. 1994). The presence of each cleavage site can serve as a character for the reconstruction of phylogeny. Abbreviations on the left side correspond to different restriction enzymes, the line represents a strand of mitochondrial DNA, all cleavage sites are numbered.

The enzyme EcoRI for example recognizes the nucleotide sequence GAATC and cuts the DNA molecule between G and A. The probability of homology of the cleavage site is much lower than the one of long primer sequences consisting of, for example, 20 nucleotides (expected frequency of analogous cleavage sites is ca. 4^{-5} instead of ca 4^{-20}). The loss of the cleavage site can occur with any mutation at one of the cleavage site positions.

More specific than the homologization of single sequences at the cleavage sites is the homologization of the *arrangement* of cleavage sites. To localize cleavage sites a cleavage site map has to be prepared (s. Fig. 107), a time consuming procedure. On this basis a character matrix can be elaborated to record which cleavage sites are present in which individuals. The following modifications of characters can occur:

– loss of the cleavage site through point mutation (GAATC becomes GGATC or GATTC etc.). As not only some identical but also different mutations produce the same loss of the cleavage site, the assumption that the loss is homologous is afflicted with a risk.
– Origin of a new cleavage site: as in the case of losses there are also several possibilities for the appearance of new cleavage sites, and so the occurrence of a new cleavage site is not a fool proof evidence for homology. However, the probability that a new cleavage site evolves from a given sequence is much smaller than the probability for loss of a cleavage site, as only a specific mutation can transform a nucleotide pattern into a recognition site (example: GGATC has to mutate to GAATC).

The probability is very low that the same cleavage site is affected just by chance in closely related species which share very similar sequences. This is so because only few mutations will be found and there exist many alternative positions for hits.

The comparison of fragment lengths (**RFLP** analysis; RFLP = Restriction Fragment Length Polymorphism) is less time consuming than the mapping of cleavage sites. The length of fragments is modified by

– insertions of nucleotides,
– deletions of nucleotides,

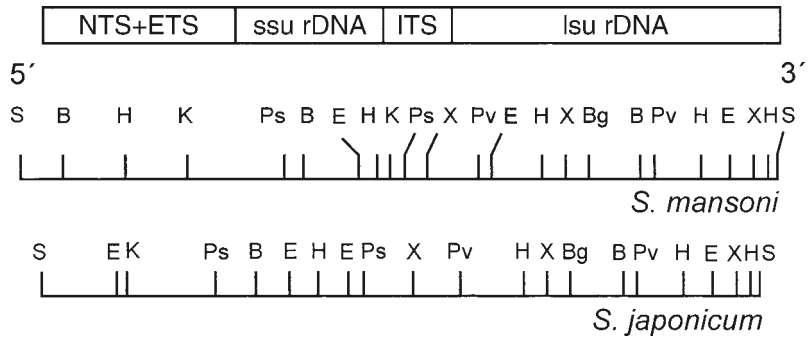


Fig. 108. Map of cleavage sites of rDNA sequences of the pathogens causing bilharzia *(Schistosoma mansoni* and *Schistosoma japonicum*). Individuals of *S. japonicum* have the same order of cleavage sites independent of their geographic origin (China, Philippines). The letters indicate for which enzyme a cleavage site is present. The bar above indicates the arrangement of sequence regions: NTS: non transcribed spacer, ETS: externally transcribed spacer, SSrDNA: rRNA gene for the small ribosomal subunit, ITS: internal transcribed spacer, LSrDNA: rRNA gene for the large ribosomal subunit (after Bowles et al. 1993).

– transpositions and inversions of sequence sections which contain cleavage sites,
– loss of a cleavage site,
– appearance of a new cleavage site.

As the loss or the emergence of a cleavage site changes the length of two fragments and also fragment number, characters are not independent. A single mutation can produce three novelties (for example one fragment loss, two new fragments). Furthermore, a certain fragment length can result by chance from mutations of different sequences. An insertion might produce a longer fragment, which by chance may match the length of another non-homologous fragment. The simple presence of a fragment of specific length cannot be a very safe argument for the establishment of a hypothesis of homology. More valuable is the generation of complex patterns consisting of many fragments comparable to a fingerprint (see criterion of complexity, ch. 5.1). When two complex restriction fragment patterns agree in detail, the probability of homology for the individual fragment is higher. The probability of homology is also higher in sequences which evolve slowly, and thus are affected by few mutations (phenomenological argument based on the evaluation of the noise-signal relation: ch. 6.5.1).

Distance values can be inferred from the presence of fragments (ch. 14.3.6, Fig. 191). If fragment lengths are to be used for a phylogenetic cladistic analysis, one should attempt to weigh the fragments according to estimated probabilities of homology. Weighting is usually justified with assumptions on the evolutionary *process*. Taking alone the fact that losses of cleavage sites at one specific site are much more probable than their appearance, the opportunity arises to weigh according to probability of events (see Albert et al. 1992). Another popular approach consists of the calculation of a distance matrix which contains values for the estimated divergence between pairs of species (ch. 14.3.6). These data can be analysed with the neighbour joining method (14.3.7).

5.2.2.5 Immunology

Comparisons of proteins with immunological methods yield data on the similarity of proteins (e.g., microcomplement fixation, precipitation tests; laboratory protocols in Maxson & Maxson 1990). Antibodies are produced against a specific reference protein, the extent of the reaction between antibody and antigen is evaluated. This reaction depends on the affinity and specificity of the antibodies and on the structure of the binding sites of the tested protein. The more similar the binding site of the tested protein is to the one of the reference protein, the stronger is the reaction. It has been shown that for limited divergence intervals between species the immunological reaction is a measure for the correspondence of amino acid sequences. Sequence similarity can be estimated with more precision the purer and

more specific the antibody and the purer the isolated test protein is.

For phylogenetic analyses, a protein is isolated from different species and each variant is compared with the one of the reference species. The antigen of the reference species is called "homologous", the one of other species "heterologous". The extent of the dissimilarity in heterologous tests is expressed relative to the homologous immunoreaction and is considered to be an indication for genetic distance. Using more reference species the possibility arises to compare distance estimates for different antigens and to determine errors in measurements. It is the goal of the experiments to obtain a matrix with pairwise distances (compare Fig. 162). The distances can then be analysed with clustering methods (compare ch. 14.3.7).

It has to be considered that the 5-10 amino acids of the antibody binding site are not representative for the whole protein and that the binding energy for the antigen/antibody association does not necessarily decrease linearly with the divergence of proteins. Two non-homologous mutations at the binding site can cause the same decrease of binding force. As in other indirect distance methods, discrete analogies and homologies cannot be distinguished. The individual evolutionary novelty cannot be identified, different non-homologous novelties cannot be distinguished when they have the same immunological effect.

Whoever works with immunological methods must not forget that with increasing distance multiple substitutions accumulate and therefore corrections are necessary (see Dayhoff et al. 1978). Multiple substitutions are chronologically succeeding mutations at the same position of the molecule. They do not modify the immunological distance and have the same effect as a single mutation. The consequence is an underestimation of the genetic distance. Furthermore, as in other distance methods for the estimation of divergence times (ch. 2.7.2.3) it has to be presupposed that substitution rates are relatively constant in time. Deviations from a regular evolutionary rate are large when comparing populations of the same species. Only with increasing divergence time a stable average for a larger period of time results. With a further increase of distances, multiple substitutions accumulate so that the estimation of divergence times becomes inaccurate. For serum albumins the detection limit is reached with divergence times of more than 120 million years (Joger 1996).

As in all methods of phylogeny inference, problems caused by analogies and plesiomorphies have to be considered (chapters 6.3.2, 6.3.3).

5.2.2.6 Homologization of isoenzymes

Isoenzymes are distinguishable enzymes with the same function but with different structure. Differences in structure can reflect origin from different gene loci (e.g., after gene duplication) or from different alleles of the same locus (allozymes = allelic isozymes). In polymeric enzymes (enzymes composed of several subunits) differences can exist in structure and number of protein monomers. For phylogenetic analyses predominantly allozymes are used, because for these orthology of coding gene loci is expected. With longer divergence times the variability can become so large that homology of novelties cannot be detected any more and backwards and parallel mutations may produce larger numbers of analogies.

Isoenzymes are separated by electrophoresis and stained by coupling the enzymatic reaction to the formation of a soluble dye (usually a formazan salt). Depending on the design of the experiment, the following assumptions are required:

- differences in the mobility of proteins in the electric field reflect differences in coding DNA,
- the enzymes of both alleles of diploid organisms are transcribed in the same amount,
- the expression of enzymes in tissue samples of different origin is comparable.

When doubts about the validity of these assumptions exist, further analyses have to be carried out to detect possible sources of error. If it is assumed that an allele is not transcribed, its expression can be tested in homozygous descendents after crossing experiments.

In the end those enzymes are homologized which catalyze the same staining reaction and travel equally far in the electrical field. The homologi-

zation of proteins of similar mobility can be uncertain because

– proteins of equal mobility can have different amino acid sequences.

With careful application of several methods of protein separation usually the larger portion of variants is discerned correctly (Murphy et al. 1990). However, the distinction of enzyme variants does not allow quantitative statements on the extent of differences in the corresponding coding DNA sequences. Proteins can only be described as being "identical" or "different".

The data gained by enzyme electrophoresis can be evaluated in different ways:

1) A data matrix can be compiled, in which the lack or presence of isoenzymes is coded for each individual of a species. This matrix corresponds to a character matrix (Fig. 122) with discrete characters.
2) A character matrix can be transformed into a distance matrix.
3) The matrix gets a third dimension when the frequencies of alleles of different populations or species are considered as well.

These data can be analysed for population genetics or for phylogenetic research. In the following only the latter will be considered.

Alleles as discrete characters

Alleles can be treated like morphological characters, when the species do not show intraspecific polymorphism and single alleles are characteristic for species or groups of species. The lack or presence of alleles are character states, character state polarity has to be established by outgroup comparison. Data can be evaluated cladistically (ch. 6) or they can be transformed into distances. The amount of differences between pairs of terminal taxa is rated as distance. A distance matrix obtained in this way can be analysed further with distance methods (ch. 8.2). However, it is doubtful whether these distances have a relation to divergence time, because the quantitative difference of the coding DNA sequences remains unknown. Furthermore, as in the case of restriction fragments, the danger exists that dependence of characters is not noted: an allele can appear as missing while it was transformed into another seemingly new one, wherefore there seem to be two events where there was only one mutation. Consequently, the genetic distance is overestimated.

The presence of specific allozymes can be used to differentiate species: Narang et al. (1989) identified in this way cryptic species of the *Drosophila quadrimaculatus* complex (Insecta: Diptera). Patton and Avise (1983) used the electrophoretically determined presence of allozymes ("electromorphs") as discrete characters for analyses of the phylogeny of waterfowl (Anatidae), and the authors tried to identify apomorphies for species groups.

Allele frequencies as characters

One has to be aware of the fact that allele frequencies are not species-specific characters, but characteristics of a population at time *t*. Frequencies change in the course of time through selection, gene flow and genetic drift *independently* of substitution rates of the DNA. Allele frequency data are better suited for population studies than for phylogenetic analyses.

Frequencies can be used as characters when the same alleles are present in several species (Swofford & Berlocher 1987). It has to be taken into account that sampling errors can be large when the allele frequencies vary intraspecifically from population to population. Whether events which lead to a shift of allele frequencies occur convergently cannot be determined *a priori*. However, there is the possibility to analyse frequencies of a large number of alleles in order to obtain a more complex pattern. The probability that changes of frequencies in the same taxa are identical by chance for many alleles is lower the more gene loci are considered. It is not possible to search for individual apomorphies, because no discrete characters are available. Allele frequencies are not used for phylogenetic studies any more. How frequency data can be converted into distances is explained in ch. 14.3.5.

5.2.2.7 Cytogenetics

For taxonomic studies occasionally changes in the number of chromosomes or of distinguishable chromosome structures are used. The comparison of numbers alone cannot serve the characterization of taxa because the homology of a number alone is dubious. From the discovery that among the primitive gastropods the Acmaeidae have 10, most Neritidae 12, and most Trochidae 18 chromosomes (Nakamura 1986), no conclusion on the phylogeny of gastropods can be deduced. To define visible changes as individual evolutionary novelties, single chromosomes or parts of chromosomes have to be homologized. A useful aid for this purpose are staining techniques that colour differentially chromosomal areas (compare methods in Macgregor & Varley 1983, Summer 1990). These techniques are not always successful and have to be tested for each case. Ideally, individual fusions or breaks can be identified. Progress in the identification of individual chromosomes and for the reconstruction of larger chromosomal rearrangements can be expected with the use of in-situ-hybridization in combination with fluorescence staining to identify the position of single groups of genes. The interpretation follows as with other discrete characters (ch. 6).

5.2.2.8 DNA-Hybridization

It is the goal of DNA-DNA-hybridization to obtain a measure of sequence differences for large portions of the genome in pairwise comparisons of species. This measure is not the number of variable sequence positions, but a distance relative to other species. The distance between two sequences increases with autapomorphies occurring in the sequence of a single organism or of the monophylum represented by the sequence. It decreases with chance similarities, symplesiomorphies and synapomorphies that are present in both sequences. Distance measurement is based on the indirect observation of the binding force between complementary DNA strands, which depends on the number of hydrogen bonds. This again is the smaller the more dissimilar the sequences are, because more unpaired positions occur. Furthermore, GC-pairings are stronger than AT-pairings, so that not only the number of paired positions influences the measurement but also the base frequencies.

Hybridization data are obtained by measuring the melting temperature (temperature at which paired strings separate) of hybrid DNA double stands. For this purpose long DNA-fragments are cut into small pieces and repetitive DNA-fragments are eliminated. Comparable melting temperatures are obtained through standardized recording of the reassociation of the DNA fragments at different temperatures. The distance to a reference species is tested. It is estimated that a difference of 1 °C in the melting temperature corresponds to about 1 % sequence difference. Hybridization can be done with DNA of the same species (control value for the reference species) or of two different species (reference species/species to be tested). The larger the measured temperature difference between intra- and interspecific hybridization, the larger is the genetic distance between a species and the reference species (further details in Werman et al. 1990). Some methodological disadvantages are the sensitivity of the measurements to variations of experimental conditions and the relative large amount of DNA required.

The following sources of error can occur:

- errors of measurement (a strict observation of standardized laboratory protocols is required),
- analogies (chance similarities in 2 species) decrease the measured distances,
- autapomorphies of a single sequence, which are irrelevant characters, increase the distances,
- when species sampling is insufficient symplesiomorphies produce wrong sistergroup relationships (see ch. 6.3.3),
- differences in mutation rates along stem lineages of terminal taxa can cause differences between measurable distances that are not proportional to the real divergence time,
- as the course of the hybridization of single-stranded fragments depends on the number of different fragments present in the solution, species with very different genome sizes cannot be compared,
- the intraspecific variability can be noticeable; for phylogenetic studies the intraspecific distances have to be distinctly below the interspecific ones.

Hybridization data are not additive (see ch. 14.3.3), wherefore simple clustering methods cannot be

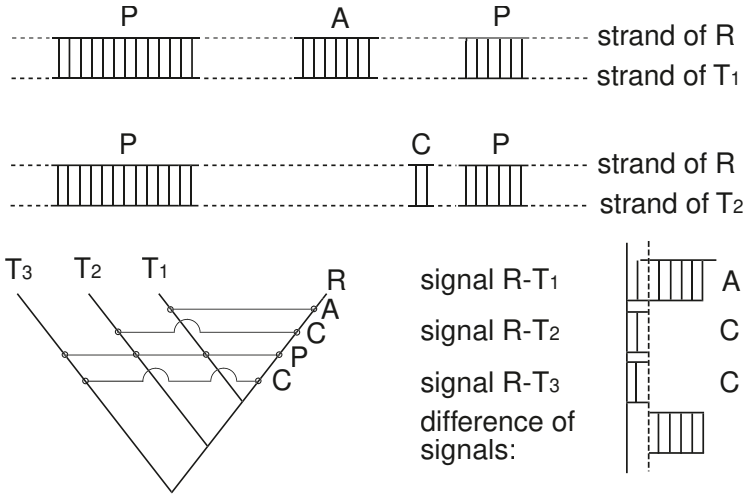


Fig. 109. Scheme to explain the mode of function of DNA-DNA-hybridizations. The sistergroup relationship between taxon T_1 and the reference species R is detected by the fact that the melting temperature of the hybrid-DNA R-T_1 is larger than for R-T_2 or R-T_3. A difference of melting temperatures ("signal difference") correlating with phylogeny is found when the number of chance similarities or convergences **C** is lower than the number of synapomorphies **A**. **P**: plesiomorphies.

used for tree construction. It has to be considered for the evaluation of this type of data that the measured distance is smaller than the real divergence, because back mutations and analogies are frequent. Therefore it is recommended to conduct corrections in the same way as in distance analyses of DNA sequences (s. ch. 8.2). Usually the simple Jukes-Cantor model (ch. 14.1.1) is sufficient to correct for chance similarities because divergence times are relatively small for most published examples (Wertman et al. 1990). Without these corrections the number of autapomorphies of terminal species is underestimated, but experience shows that the topology of dendrograms does not change much with more sophisticated corrections. As in contrast to sequencing of single genes the number of nucleotides that are considered in DNA-DNA hybridizations is very high, it is expected that deviations from the expected number of analogies occurring on average is small when the model chosen for corrections simulates the real sequence evolution. False signals due to chance similarities (noise) should compensate each other while true homology signals should increase on average linearly with increasing number of nucleotides (Fig. 109; for additivity of signal see also Fig. 154).

Hybridization data do not allow the identification of taxon-specific characters, evolutionary novelties are not identified individually. With DNA-DNA-hybridizations single nucleotides, insertions and other mutations are not considered, but a multitude of genes and other sequences, which together contain a very complex pattern that is characteristic of a species. These patterns themselves are not visible and cannot be described, but rather the extent of the differences between two patterns is indirectly measured and quantified. The complexity of potential synapomorphies shared by two taxa (the "signal" in favour of a monophylum) is recorded with a unit of measurement, and usually it is not attempted to estimate from this the number of shared substitutions. False signals (noise) result from analogies. The real "signal" has to be stronger than the "background noise" to be detectable.

The most famous phylogenetic analysis performed with this method is the bird phylogeny analysis by Sibley & Ahlquist (1990). For example, they obtained evidence that the enigmatic South American hoatzin (Opisthocomidae; see also Fig. 113) belongs to the cuckoos and allies (Cuculiformes), that owls are not related to diurnal birds of prey but to podargids and goatsuckers (Caprimulgiformes), that the vultures of South

America (Fig. 70) are not related to the old world vultures but to the storks (Ciconiiformes); the American blackbirds (Icteridae) are not related to ravens (Corvidae) but the sister taxon to the new world mocking birds (Mimidae).

5.2.2.9 RAPD and AFLP

The **RAPD method** ("random amplified polymorphic DNA", Williams et al. 1990) consists of amplifying unknown DNA regions of about 200 to 2000 bp length with the help of random primer sequences using the PCR-technique. It is not of interest from which sequence areas or genes the sequences are amplified, the loci remain "anonymous". To amplify only regions which are limited by sequences that are exactly complementary to the selected primers, the experimental conditions of the PCR reaction have to be adjusted in such a way that unspecific primer annealing is avoided. Unspecific amplicons would give false signals. An electrophoresis gel with a typical RAPD product contains several bands of DNA fragments, the RAPD markers, which were amplified from the complete DNA of an organism. The pattern of fragments is used as "fingerprint". The laboratory methods are not complicated, it is not required to design specific primer sequences as for the amplification of specific genes. Problems can arise with the reproducibility of single experiments, which depends on the reaction parameters. An advantage of the method is the relative small laboratory expenditure. Users of this method can find recipes and recommendations in the literature (e.g., in Berg et al. 1994, Grosberg et al. 1996, Burwo et al. 1996, Benecke 1998).

Often polymorphisms are detected when individuals of different populations or species are compared, i.e. variations of the presence of bands on the gel, which are caused by evolutionary elongation or shortening of sequences or by loss or emergence of a primer-recognition site. Such variations can be used to identify populations or species. To detect diagnostic bands one usually has to try several randomly chosen primers. In this way, for example, the morphologically poorly distinguishable species of malaria-transmitting mosquitoes (*Anopheles* species) can be clearly identified (Wilkerson et al. 1993). Also population studies can be done with RAPD analyses.

Hypotheses of homology for amplified sequence fragments of equal length are based on the assumption that it is little probable that randomly amplified regions have the same length in two species by chance or that different primers amplify the same homologous section. Thus a high probability of homology for RAPD fragments is taken for granted. These hypotheses can be tested by sequencing the fractions. However, this effort is usually not realized. For phylogenetic studies, the presence of a band can be coded like a character state to construct dendrograms with cladistic methods (ch. 6.1.2).

The **AFLP technique** (Zabeau & Vos 1993, Sharbel 1999) combines PCR with RFLP (restriction fragment length polymorphism). The method is more stringent than RAPD and repeatable. The major steps are: (1) digestion of genomic DNA with restriction enzymes, (2) ligation of restriction fragments to oligonucleotide adapters, (3) amplification of a subset of ligated fragments via PCR, (4) fragment separation by gel electrophoresis. Usually, DNA is digested with two restriction enzymes, one which cuts often (4 bp-recognition sites) and the other less frequently (6 bp-recognition sites). Only those fragments of medium length (50-600 bp) and with two different cleavage sites are subsequently amplified. The synthesized adapter to which the fragments are ligated consist of a core sequence complementary to artificial PCR primer sequences and a restriction recognition sequence. Primer sequences are designed to get stringent PCR conditions. For the first amplification round primers are used that are complementary to the sequence of the adapter plus the restriction sites, and a single nucleotide is added at the 3'end that will match to only part of the fragments. The amount of amplified fragments is thus reduced (preselection). The second amplification is done with primers identical to those used for the first PCR round, however, with one or two additional nucleotides at the 3' end to increase selectivity. Presence and absence of fragments is compared assuming that fragments of the same length are homologous. Fragments of the same length have the same restriction sites at both ends and in addition the two to three nucleotides used for fragment selection during PCR. Gain of a particular AFLP locus is much less probable than loss, because losses can result from several non-homologous mutations (mutations at different positions of the

restriction sites, mutations at the positions of the selective PCR sites, insertions or deletions within the fragment region).

The AFLP technique is useful to study polymorphisms within species, some loci may be so variable that they can be used to discern individuals. AFLP bands coded as presence absence data are dominant markers. It is possible to discover markers characteristic for clones, populations, species, or groups of closely related species. If required for population analyses, the genotype of a marker can be inferred from the optical density of a fluorescent band on the electrophoresis gel: the fluorescence is expected to be larger for homozygous than for heterozygous individuals.

5.2.2.10 Amino acid sequences

Sequences of amino acids are often not obtained by sequencing of proteins, but translating coding DNA sequences, which is easily done with computer programs like MacClade (Maddison & Maddison 1992). Principally, amino acid sequences can be analysed in the same way as DNA sequences. Differences exist in the weighting schemes for substitutions. Weighting is usually done with empirical data which reflect the estimated probability of events (i.e., the probability for specific substitution events):

To use empirically determined substitution frequencies a matrix is constructed which contains a probability value for each pair of amino acids (Dayhoff et al. 1978). This value is an estimation of the probability that a specific amino acid is substituted by another one. A diagonally symmetrical matrix with 400 entries (20 times 20 amino acid substitutions) has been used as a universal model for substitution probabilities. An example is the Dayhoff matrix. For its compilation 1572 mutations were analysed in 71 sequence families of which the phylogeny had been reconstructed. In each dataset the frequency of amino acids was counted to estimate the probability for an exchange. The units used for the matrix are called PAM units ("point accepted mutations per 100 residues per 10^8 modelled evolutionary years"). In this matrix (appendix 14.11) those pairs have a negative value which empirically occur less frequently than expected from a random combination of amino acids. Positive numbers characterize pairs which were observed more often than expected. The cause for the empirically observed variability in substitutions lies in the physical and chemical properties of the amino acids. The application of the Dayhoff matrix requires the assumption that substitution rates for individual amino acids are different, but rates neither depend on sequence position nor on the examined lineage of organisms.

For the evaluation of chemical similarity of amino acids, the consideration is important that enzymatic activities depend on the physical-chemical properties of the amino acids and therefore any variation of a protein is under different selection pressure depending on the function of the mutated site and the properties of the new amino acids. Amino acids are classified according to molecule size, hydrophobicity and polarity (McLachlan 1972, Taylor 1986a,b).

As analyses at DNA level are much more popular, further details of the interpretation of peptides are omitted here.

5.3 Determination of character polarity

It is elucidated in chapter 3.2 that a tree in which groups of taxa are distinguished can be constructed without knowledge of character polarity. But such a graph is unrooted and not polarized, and assuming that the topology is correct the groups separated by edges can be either monophyla or paraphyla, the direction of evolution is not depicted. Therefore, a directed or rooted dendrogram is more informative and necessary when phylogeny is to be illustrated. We have to distinguish between the **polarity of characters** and the **polarity of the tree**. By convention, in cladistics "polarity" usually refers to characters. The determination of the polarity of a dendrogram is also figuratively called "rooting". The root is the basal point of origin of a phylogenetic tree.

It can be attempted to determine the chronological order of character emergence or of character transformation for all characters which are assumed to be homologies. This concerns (a) a character present in some taxa and completely lacking in others, or (b) two or more character states of a frame homology. Since it is possible to define the absence of a character as a state, or to describe the different states as characters (= detail homologies), there is principally no difference between case (a) and case (b) except that the states may get different weights.

The assumption that the more frequent character state is the more primitive one is undoubtedly useless: the ontogeny of most tetrapods, for example, does not include aquatic larvae, nevertheless the assumption that the life cycle of amphibians is more primitive is well founded. To assume that ancestral characters are found in populations at the center of origin of a radiation while at the borders of the area of distribution more derived characters are found might be correct in many cases (chorological criterion). However, this criterion is not reliable, because evolution continues also in the areas of origin. For the determination of polarity the following approaches are popular:

a) phylogenetic character analyses with outgroup comparison (ch. 5.3.2),
b) cladistic outgroup addition ("outgroup comparison" of phenetic cladistics, see ch. 5.3.3, 6.1),
c) estimation of the evolutionary increase of complexity (ch. 5.3.4),
d) the ontogenetic criterion (ch. 5.3.5),
e) the paleontological criterion (ch. 5.3.6).

These methods are introduced in the following.

5.3.1 Ingroup and outgroup

In practice the **ingroup** is nothing else but a selected group of species whose monophyly is established or to be tested. The **outgroup** is the group of all organisms which do not belong to the ingroup, and thus includes "all other living organisms". In phylogenetic cladistics, the members of the ingroup can be identified with constitutive characters (see ch. 4.2.2) regardless whether the group is monophyletic or not*. All organisms, which do not show these properties do not belong to the examined ingroup. This is exactly the way characters are analysed with Hennig's method (ch. 6.2): searching for analogies or homologies, all known organisms are considered. During the search for homologous characters the ingroup can be extended as soon as more taxa are found which show the same constitutive characters. This is a **phylogenetic character analysis with outgroup comparison**.

The method of phenetic cladistics as used by many systematists (ch. 6.1) does not require the consideration of all available knowledge on other organisms and is not based on character analyses. This is partly a consequence of the practice to leave the search for homologies to a computer system, which always is fed with only limited data on a restricted number of organisms, and is also part of the logics of phenetic cladistics which implies that homologies are identified *a posteriori*. This restriction is a source of errors, because ple-

* A group is not monophyletic when it is composed of those species that show an apomorphy of the taxon's ground pattern, excluding species where the apomorphy is secondarily substituted.

siomorphies can be mistaken for apomorphies (see plesiomorphy trap, ch. 6.3.3). The algorithms of cladistics do not require that more than one taxon is defined as "outgroup". For this reason two different forms of outgroup comparison have to be distinguished.

Outgroup character comparison: comparison of characters to estimate the probability that a character or character state is an evolutionary novelty (= apomorphy) of an ingroup and that the absence of this character in all other organisms (= the outgroup) is the phylogenetically older state (= plesiomorphy).

Cladistic outgroup addition: determination of at least one taxon as outgroup for rooting of a dendrogram. A character analysis does not take place.

5.3.2 Phylogenetic character analysis with outgroup comparison, reconstruction of ground patterns

When we examine the phylogeny of a selected group of organisms it is of interest to distinguish different monophyla within this group. For each putative monophylum a character analysis has to be performed to identify the respective autapomorphies and to work out a ground pattern for each monophylum. The **ground pattern** represents a reconstruction of elements that must have been present in a last common ancestor. Each putative monophylum is defined as a functional ingroup for the corresponding character analysis. The phylogenetic character analysis of an ingroup consists of the following steps:

- choose a group of species for which monophyly has been corroborated. When such a group of species is not known, start with individual species and look for the next similar ones. Define this group of species as the ingroup (in any case the working hypothesis is that the ingroup is a putative monophylum).
- Search for identities or similarities which occur in all or in some of the species of the ingroup. Analyse each of these characters individually.
- Choose only those characters which occur in all species of the ingroup or for which it can be shown that they are secondarily reduced or modified in cases where they are missing. The selected characters can be called potential **characters of the ground pattern** of the ingroup (reconstruction of the ground pattern: see below). The further analysis can only be performed with characters of the ground pattern.
- Describe each character in as much detail as possible. The more details that are known to be present in all species, including the contacts to neighbouring characters in space or time, the greater is the probability that the character is homologous (see ch. 5.1).
- Choose only those characters of the ground pattern for the further analysis for which the estimated probability of homology is high.
- Check whether a character also occurs outside the selected group of species. To do so consider all known organisms, scrutinizing especially those that are most similar and seemingly closely related, but not only these.
- If a character is present also outside the ingroup then it is a **plesiomorphy** of the ingroup's ground pattern. The hypothesis of plesiomorphy can be weakened by evidence that suggests analogy, **convergence or parallelism**.
- A character that is only present in the ingroup can be called a potential evolutionary novelty (**autapomorphy**) of the ingroup's ground pattern. This argument holds only with characters of high probability of homology, because otherwise the correspondence could as well be a chance analogy. (Attention: if a character has more than two states, for example state 0 and 2 in outgroup taxa, state 1 in the ingroup, it must be examined if 2 might be a derived form of 1. In that case 1 is not an apomorphy.)
- Whenever an autapomorphy has been identified it serves simultaneously and inevitably as evidence in favour of the hypothesis of monophyly of the ingroup.
- The ground pattern of the ingroup consists of all identified plesiomorphies and autapomorphies.

Evidence that similarities could be **convergences** is obtained by analysing the details and the variability of a character. Superficial correspondence and many differences in detail, or possibly even a high intraspecific variability of a character (e.g., number of hairs, form of pigment spots) increase the probability that corresponding properties can

originate by chance. It is also possible that two structures which are similar at first sight prove to be homologous to different originals when examined in detail. A major problem is that differences alone are not evidence for non-homology, simply because "copies can become very noisy" (ch. 4.3.1). Additional information is required, ideally incompatibility with the distribution of other characters of high probability of homology. Putative convergences and analogies can be recognized *a posteriori* after the reconstruction of the dendrogram, where they appear as homoplasies (compare chapters 4.2.1, 4.2.2).

In this way a ground pattern is reconstructed character by character and at the same time the hypothesis of monophyly is substantiated. In the further course of the analysis of phylogenetic relationships of higher ranking taxa the monophylum takes the role of a terminal taxon and has to be represented by characters of its ground pattern.

An apomorphy can also be considered as character state of a larger frame character. Working with "character states" implies that there exists a complex frame homology in which detail homologies that are evolutionary novelties or plesiomorphies can be discerned. Homology of the frame should be corroborated, its details are relevant for the analysis of polarity described above. A frame lacking a novelty has a plesiomorphic state, with the presence of a novelty it is called an apomorphic character. The differentiation of these states or the distinction between presence and primary absence of evolutionary novelties is the "determination of polarity".

A more formal description of the argument is: the frame homology ***R*** occurs in species of the outgroup $\boldsymbol{T_A}$ and in the ground pattern of the ingroup $\boldsymbol{T_I}$, the outgroup not being a monophylum. A detail homology ***D*** of ***R*** only occurs in the ground pattern of $\boldsymbol{T_I}$. Conclusion: ***D*** is a putative evolutionary novelty of $\boldsymbol{T_I}$ supporting the monophyly of $\boldsymbol{T_I}$ (explanation: the outgroup should not be a monophylum because otherwise the danger exists that a plesiomorphic character state of the ingroup is erroneously considered to be apomorphic when this state does not occur in the sister taxon; the apomorphic state could be that of the monophyletic outgroup).

The following analysis serves to test the hypotheses of monophyly and homology that have been found with the previous steps:

- search for further putative apomorphies for the monophylum that is being studied,
- search for other homologies and test whether the distribution of characters on species and groups of species is compatible with the hypothesis of monophyly (see Figs. 78, 122, 193, 201).
- The more potential apomorphies of high probability of homology support the same group, the higher is the probability that (a) the individual characters are evolutionary novelties and (b) the group is monophyletic. When the species groups incompatible to this group are supported with markedly fewer apomorphies of similar quality, the characters in question are probably analogies. The test for compatibility of all discrete characters of a dataset is usually done with a maximum parsimony analysis with weighted characters (ch. 6.1.2).

The outgroup comparison of this analysis consists of a search for characters of the ingroup species in all other organisms (the outgroup) to identify a potential evolutionary novelty.

For clarity, when coding polarized characters for cladistic analyses all plesiomorphic characters should be marked with a "0" in the data matrix (ch. 6.1.1), all apomorphies (when only two character states are present) with a "1". The proposed polarity can be enforced in MP-analysis (ch. 6.1.2) by defining a hypothetical taxon as outgroup for which all characters have the plesiomorphic state ("all-zero-ancestor"). This represents the hypothesis that a last common ancestor with these character states existed. Such a coding should not be based on *ad hoc* assumptions but on a well-founded *a priori* character analyses.

Example: Within the Cnidaria, the Hydrozoa, Cubozoa and Scyphozoa have similar cnidocytes with a stiff cnidocil. These taxa could be selected as a functional ingroup with the name "Tesserazoa", all other organisms form the outgroup. A cnidocil (the cilium that causes the discharge of a stinging cell) and a cnidocyte occur outside these groups only in the Anthozoa. Further analysis of the cnidocil shows that (as far as known, see Ax 1995) in the ingroup the cnidocil does not

show distally the usual 9×2+2 pattern of microtubules but a variable number of single microtubules. Furthermore, ciliary rootlets and accessory centrioles are missing in comparison with cilia in outgroups. The cnidocil and the corresponding cnidocytes can be regarded as frame homology, the arrangement of microtubules as well as the lack of ciliary rootlets and centrioles are detail homologies. The modifications in comparison to the outgroup are potential evolutionary novelties. The comparison shows that ciliary rootlets and accessory centrioles are common in the animal kingdom and also occur in the Anthozoa, while the considered frame homology (cnidocytes) only occurs in Anthozoa and Tesserazoa. Taking the stiff cnidocil of the Tesserazoa as an apomorphic character state, the complete apomorphic pattern is more complex than, for example, the character "accessory centriole reduced" due to the simultaneous occurrence of several potential detail homologies, and it has a higher probability of homology than for a single subordinate detail homology. The plesiomorphic character state is therefore the one found in Anthozoa. This conclusion is enhanced by the fact that plesiomorphic details of the fine structure of the cilium of Anthozoa are common in the outgroup.

In this example the "determination of polarity" distinguishes plesiomorphic and apomorphic character states of the cnidocil. Further argumentation requires a search for additional characters which support the split {Tesserazoa, outgroup} and the hypothesis of homology for the abovementioned novelties of the Tesserazoa. Such characters are the presence of cnidocysts of the type "microbasic eurytele" in the ingroup and of a linear mtDNA molecule which is usually circular in the outgroup and especially in the Anthozoa. An additional correspondence is the presence of medusae in all ingroup taxa. The examination of the probability of homology of the character "medusa present" raises doubts, because the medusae have a different anatomy and develop in different ways. Known facts could as well be explained with convergent evolution of medusae in Hydrozoa and Scyphozoa + Cubozoa. Therefore this character is not used to support the hypothesis of monophyly of the group "Tesserazoa" (Ax 1995).

Reconstruction of ground patterns

A ground pattern is the sum of all characters assumed to be present in *ancestor individuals* or in an *ancestral population*. It is a result of character analyses or of the comparison of a tree topology with the distribution of character states in terminal taxa, whereas a "bauplan" is a sort of "general anatomy" with an arbitrarily selected collection of characters. The ancestral population is the last one before the speciation that gave rise to the first split in the monophylum of the corresponding ground pattern and it therefore is the stem-population of this monophylum (compare ch. 2.6). Considering long periods of time and large groups of species, one also talks of characters in the ground pattern of the *last common stem species*, or, less precise, of the *last common ancestor*, without distinction of different populations of this species, because in practice mostly no information on the evolution of the stem species is available. The ground pattern contains old characters, plesiomorphies, and evolutionary novelties, which evolved in the stem lineage and are autapomorphies of the monophylum. The reconstruction of these ground patterns is especially important for each terminal taxon used in a phylogenetic analysis, because terminal taxa should only be represented by ground pattern characters in a data matrix to avoid hidden sources of error.

The following steps are necessary:

In order to reconstruct the ground pattern of a selected monophylum one has to consider homologous characters of all members of the monophylum. In principle, ground pattern reconstruction by character analysis can be performed in top-down direction, starting with the smallest terminal units (usually species). It is not necessary to know details of the corresponding phylogenetic tree, but a hypothesis is required concerning the composition of the monophylum that is being examined. If the terminal taxa are species, one examines the characters that are typical for these species. If the terminal taxa are larger units, one has to work with previously reconstructed ground patterns of these taxa. Each character has to be analysed separately (Fig. 110):

1. Homologous characters occurring with the same state in all members of the monophylum are characters of the ground pattern of the monophylum. These conserved charac-

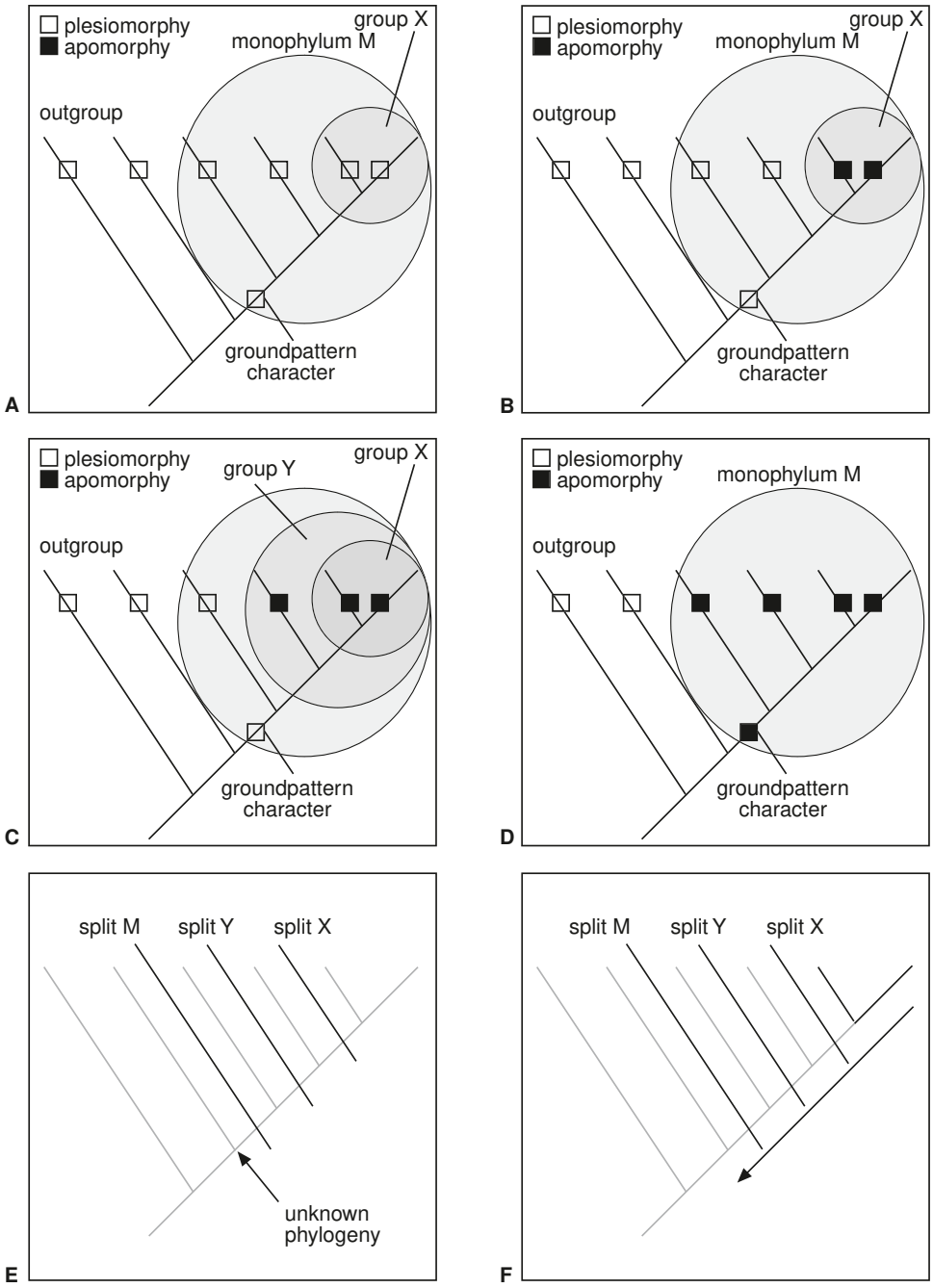


Fig. 110. Reconstruction of ground pattern characters. **A.** Plesiomorphy of the monophylum in the ground pattern. **B.** Apomorphy of group X is not an element of the ground pattern. **C.** Plesiomorphy of X in comparison to group Y is not an element of the ground pattern because the state is an apomorphy of group Y. **D.** The apomorphy of the monophylum as element of its ground pattern. **E.** In practice, single species or a group of species are compared with the rest of the species set, a split separates functional in- and outgroup. Knowledge of the branching order on the phylogenetic tree is not required, but it accumulates in the course of the analysis. **F.** Character analysis and descending reconstruction of the tree can be done simultaneously.

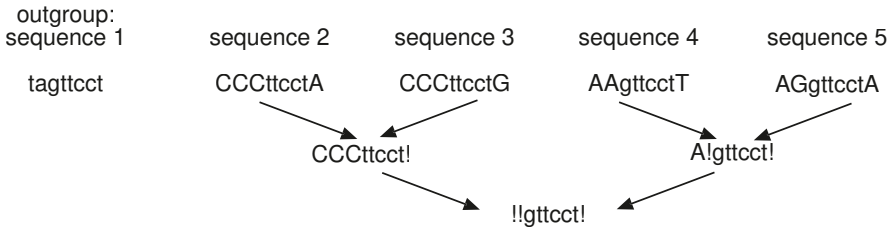


Fig. 111. Reconstruction of ground pattern character states in the case of sequences. Nucleotides in variable positions are written in capital letters, unknown character states in the ground pattern are marked with exclamation marks (!). The conserved motive "ttcct" characterizes invariable positions and most probably represents a homology of all sequences. The "g" in front of this signature occurs in a subgroup of the ingroup and in outgroup sequences (not shown) and therefore probably belongs to the ground pattern of the ingroup.

ters can be plesiomorphies (Fig. 110A) or apomorphies (Fig. 110D) of the monophylum under consideration.

2. A character analysis has to be conducted for characters with several states. For this purpose groups of species with the same character state are temporarily combined as functional ingroup and for this (or for a single species in case only one shows this state) it is tested whether the state can be an apomorphy (of the individual species or the group of species) using the procedure for character polarization (see above). This requires an outgroup comparison with consideration of all species which do not belong to the functional ingroup and to the monophylum. Apomorphies of subgroups of the monophylum do not belong to the ground pattern of the monophylum.
3. The plesiomorphic character state which is the counterpart of the apomorphy identified in the previous step has to be analysed in all other species of the monophylum. It could be that this state is plesiomorphic at the level of the previous step but it may be simultaneously an apomorphy of a larger higher ranking group within the monophylum (Fig. 110C).
4. Of the variable characters only those states are incorporated into the ground pattern which do not occur as apomorphies of subgroups in the monophylum. Highly variable characters in which back mutations and analogies occur frequently should not be considered because their evaluation is problematic (see e.g., "noise" in DNA sequences, ch. 6.5). A ground pattern should only contain character states whose polarity and homology could be postulated with some certainty.
5. In some cases it can be shown that some characters were polymorphic in the last ancestral population of the monophylum, implying that several character states may be present simultaneously in the ground pattern. Polymorphic characters are only useful for phylogeny inference when exact knowledge of their evolution is available (ch. 1.3.7).

Fig. 110 may suggest that for character analyses a phylogenetic tree already has to be known. This, however, is not correct. In practice only single splits (Fig. 110E) are considered which separate species groups in the set of taxa, the groups are used as functional in- and outgroups and examined as *potential* monophyla. For those ingroups for which apomorphies are discovered in the course of character analysis, the hypothesis of monophyly can be maintained. In this way monophyla are identified step by step. Character analysis can lead to a descending top-down reconstruction of phylogeny, but often monophyly is substantiated only for part of the species groups, and the order in which these monophyla are detected can depend on the order of characters selection (we can, for instance, examine the set of taxa that possess compound eyes, and then those that possess mandibles). But it is decisive that for each terminal taxon the ground pattern is worked out in a preceding character analysis.

Phylogenetic character analysis is more reliable in comparison to the phenetic-cladistic analysis (ch. 6.1):

– Ground patterns are reconstructed for supraspecific taxa, reducing the risk that autapomorphies of derived species are taken for characters of the whole taxon.

– The characters of all known organisms are taken into account. The cladistic outgroup comparison (ch. 5.3.3), for example, entails the danger that autapomorphies of a monophylum that has been chosen as outgroup are misinterpreted as plesiomorphies in comparison with ingroup character states.
– Characters of low probability of homology are eliminated from the analysis.

When analysing DNA sequences, ground patterns can also be reconstructed in the same way as for morphological characters. This procedure has not been used so far. However, it is convenient to represent a larger taxon by a single ground pattern sequence, especially when datasets are very large and calculations time-consuming, and to avoid errors caused by chance similarities of autapomorphic states of terminal taxa. Until recently in many phylogenetic analyses for example the sequence of a chicken has been used as if it could represent the ground pattern of birds (Aves), without testing which characters are autapomorphies of the genus *Gallus* and therefore do not belong to the ground pattern of Aves. A descending reconstruction of ground patterns could be performed as shown in Fig. 111.

Fig. 111 illustrates the fact that a sequence of a ground pattern contains less known characters than a terminal sequence, because in a phenomenological phylogenetic analysis each sequence position with an uncertain character state is eliminated. The supraspecific taxon is represented by characters which are not autapomorphies of lower ranking groups.

The cladistic reconstruction of ground patterns is further explained in ch. 6.1.2.1 (Figs. 127, 131).

Attention: the reconstruction of a ground pattern requires a **suitable sample of species**! The more taxa of a monophylum are considered the smaller is the danger that mistakes occur. The danger that autapomorphies of subordinate groups of species are included in the ground pattern exists when primitive species, or species which are the only ones that still show specific plesiomorphies, are not considered. Example: the incorrect statement that mammals are primarily (thus in the ground pattern) viviparous, can only be maintained by someone who does not know anything about the existence of the egg-laying Monotremata. In principle, the same source of errors exists in molecular systematics.

5.3.3 Cladistic outgroup addition

The cladistic determination of character state polarity requires "**rooting**" of a tree graph (a dendrogram). Rooting of the dendrogram is performed by selection of one or more outgroup taxa (Maddison et al. 1984). The assessment of character polarity is not the result of individual and *a priori* character analysis. Many cladists (users of phenetic cladistics, pattern cladists, see ch. 6.1) consider the results of *a priori* character analysis (ch. 5.3.2) to be subjective, based on unfounded 'ad hoc' decisions.

Example for mistakes that occur by cladistic outgroup addition: when one adds to a data matrix with morphological characters of isopod crustaceans (Isopoda) the ground pattern of the sister taxon (Tanaidacea, tanaids) and defines this as outgroup, one obtains a monophylum that includes all taxa having fan-shaped uropods within isopods, because in the Tanaidacea and in several groups of isopods the uropods are rod-like or lamellate with varying shapes. The fan-shaped uropods appear to be an apomorphy within Isopoda. If, however, the Mysida (opossum shrimps) are selected as outgroup, the fan-shaped uropods appear as plesiomorphic within the isopods, because the Mysida also have these fan-shaped uropods. The phylogenetic analysis requires *a priori* a detailed anatomical analysis of the homology of the different forms of uropods, especially in respect to the question whether styliform or fan-shaped uropods are convergences or homologies.

Following the selection of an outgroup, in phenetic cladistics all characters of the ingroup which have another state than in the outgroup are automatically without preceding character analysis identified as apomorphies. This is an important aspect of the cladistic homologization of characters or characters states (see also ch. 6.1, 6.1.10). For the cladistic determination of character state polarity it has only to be known in which taxa the character occurs and which taxon shall be considered as outgroup. There is no input of further information for the cladistic analysis. In contrast, the phylogenetic character analysis additionally

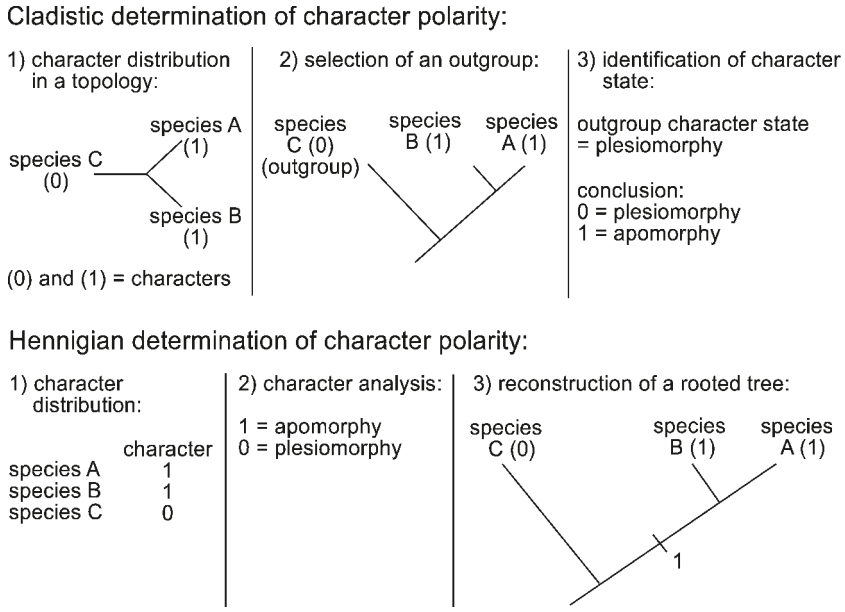


Fig. 112. Determination of character state polarity: *a posteriori* in the phenetic cladistic way and *a priori* through phenomenological character analysis. Character analysis for the phylogenetic determination of character state polarity is discussed in ch. 5.3.2.

considers the probability of homology of characters and of character states, mainly considering the complexity of their structure, and also the character distribution in *all* known taxa or in all taxa that are relevant for the analysis (see above). Experience teaches that the phenetic cladistic method leads to wrong decisions concerning character state homology and polarity, and thus to the reconstruction of unreliable phylogenetic trees.

The difference between these two methods is illustrated in Fig. 112.

Attention: the cladistic determination of character state polarity only provides reliable results when outgroup characters each show the plesiomorphic state. The following mistakes can occur unnoticed:

- An autapomorphy of the outgroup is misinterpreted to be the plesiomorphic state, the plesiomorphic character state of the ingroup is misinterpreted as apomorphic.
- Chance similarities shared by species of the outgroup and species of the ingroup are erroneously coded as symplesiomorphies (= shared plesiomorphies)

The more outgroup species are considered, the smaller is the danger that character states are interpreted erroneously. If characters are weighted according to their probability of homology and the more complex the homologies are, the smaller is the danger that chance similarities (analogies) are taken for characters of the ground pattern.

The results of a determination of character state polarity that has been performed and justified prior to tree construction can be considered in a cladistic analysis by adding a hypothetical outgroup taxon that is exclusively composed of plesiomorphies with regard to ingroup taxa character states (an "all-zero ancestor" if plesiomorphies are coded with a zero). This artificial taxon would represent the reconstructed ground pattern of the ingroup (see previous chapter).

5.3.4 Increase of complexity

As evolution of complexity requires time, one might deduce that in an evolutionary series of variations of a homologous morphological structure the simpler constructions are the phyloge-

netically older ones. This assumption is often true: the lens eye of cephalopods evolved from a pinhole-camera eye, the differentiated mouthparts of crustaceans originated from 3 pairs of nearly identical walking legs, the double circulatory system of mammals with its chambered heart is a more recent variation of the simpler system with a one-way-heart of the primarily aquatic vertebrates. The increase of complexity is also observed at the level of molecules: the 18S rDNA gene of plants and animals is in large parts homologous to the 16S rDNA gene of prokaryotes, but it carries insertions, some of which are characteristic for large species groups (Euglenozoa: elongation in the helical regions E21-9 and E21-3; Eukaryota: elongation E10-1 and E10-2; see Wolters 1991). Gene duplications at first produce redundant sequences, which subsequently can differentiate independently, with the result that duplicate genes may take up new functions (ch. 5.2.2.3). Gene duplication is undoubtedly an important evolutionary mechanism, which produced whole gene families (collagens, actins, immunoglobulins, tRNAs, globins) and increases the complexity of organisms.

The assumption that complexity increases, however, cannot be applied to every character: a snake has a "simpler" locomotion apparatus compared to other tetrapod amniotes and a reduced number of different anatomical structures. Nevertheless the method of movement of snakes is not the primitive one of tetrapods. – The parasitic Rhizocephala are crustaceans of which the females lack a gut and appendages. Their morphology shows few structures in comparison to other crustaceans: the complexity is reduced as an adaptive consequence of a parasitic mode of life. These examples show that an increase of complexity is not suitable as a criterion for the determination of character state polarity.

5.3.5 The ontogenetic criterion

For the evaluation of character state polarity it is often recommended to analyse the course of the ontogenetic development of characters. One starts with the observation that, logically, in phylogeny plesiomorphic character states occur chronologically prior to apomorphic alternatives, and one assumes that during ontogeny a similar order of character formation is maintained. However, the relevance of this rule for practical work is limited mainly due to lack of empirical data.

Biogenetic rule

This criterion can be traced back to Ernst Haeckel's biogenetic law (= rule of recapitulation): "ontogeny is a short and quick recapitulation of phylogeny …" (Haeckel 1866). Hence an early embryonic character state can be considered more primitive than a later state of the same organ (ontogenetic character precedence). Recapitulated characters are called **palingeneses**. Today it is known that recapitulation occurs in many organisms (Fig. 113), because developmental constraints conserve the anlage of embryonic or larval structures. However, recapitulation is not a law of nature valid for all characters and all organisms.

The application of the biogenetic rule for the determination of character state polarity requires the following statements:

1) A homologous adult character ***M*** is present in the species or group of species ***A*** in the state M_1, in the species ***B*** in state M_2.
2) Developmental stages of species ***B*** temporarily show the character state M_1.
3) Therefore it can be concluded that M_1 is the phylogenetically older state.

However, the character M_1 can also be a new acquisition of the larvae. This is the evolutionary process of **caenogenesis**, resulting in a **caenogenetic character**, for example the prehensile labium of the larvae of dragonflies. In this case the argumentation (Fig. 114) has to be modified: the adult character would have to be replaced by a functioning larval organ, preceding developmental stages would have to show a corresponding anlage of the organ.

Example for the argumentation:

1) Gill slits are an adult character of cartilaginous and bony fish, but are missing in adult birds.
2) Anlagen of gill slits occur temporarily in embryos of birds.
3) The presence of gill slits in adults is the phylogenetically older state.

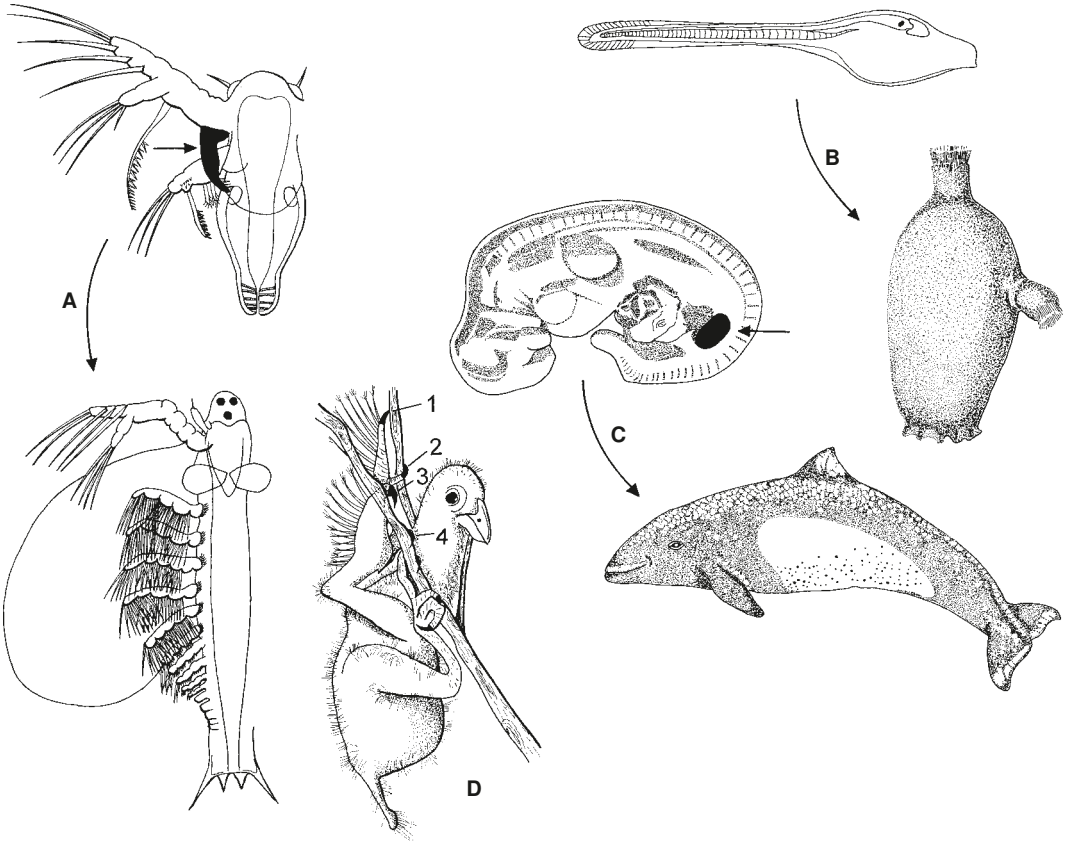


Fig. 113. Examples for recapitulations. **A.** Nauplius-larvae of many crustaceans (here *Limnadia stanleyana*, after Anderson 1967) use the second antennae as a mouthpart; the proximal endite (black), which is important for larval feeding, is reduced only in the course of ontogeny and is absent in adults. The phylogenetically older state of the first postantennal appendage is recapitulated; in the stem lineage of the Mandibulata this appendage was used to handle food in a way that is probably comparable to the function of the mandible of many modern crustaceans. **B.** The larvae of ascidians have a way of locomotion and an appearance similar to other primitive chordates. **C.** In the embryo of whales the anlage of the posterior legs is still present. **D.** Chicken of the hoatzin (*Opisthocomus hoazin*, after Attenborough 1998) still show a pair of claws on their wings, which are used to climb vegetation.

But this argumentation can also lead to misinterpretations:

1) Adult salamanders (*Proteus anguinus*, *Ambystoma mexicanum*, *Typhlomolge rathbuni*, and others) can have external gills, in most species they are missing.
2) Larvae of salamanders usually have temporary gills,
3) consequently the possession of gills in adults is the primitive state (*wrong conclusion*!).

The biogenetic rule relies on the assumption that the evolutionary addition of structures is repeated during ontogeny in the same anatomical surroundings within the organism and in the same chronological order (case of *terminal addition*). If addition of apomorphic details during phylogeny is common and reversals or deletions of derived states is rare, the ontogenetic criterion should be reliable (Meier 1997). However, in nature there exist several deviations from Haeckel's rule (see Osche 1985):

– Heterotopies are modifications of the position of the primordium of an organ: in most Malacostraca epipodial gills insert laterally on the basal segments of the thoracal appendages, while in the Amphipoda (beach hoppers) they insert medially.

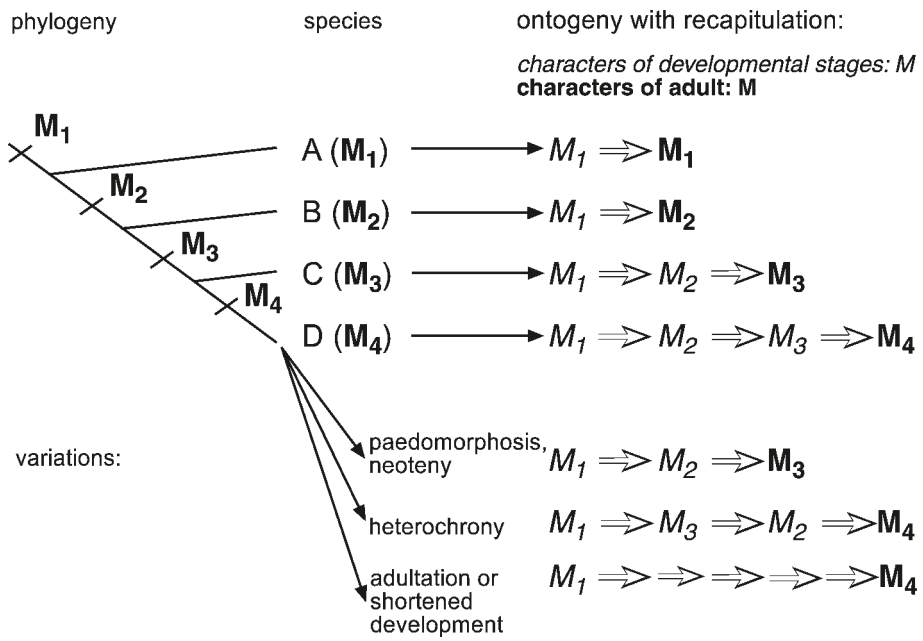


Fig. 114. Possible variations of recapitulation.

- Heterochronies: change of the time of appearance of a primordium during embryonic development: some ascidians already have siphons and some gill slits in a larval stage, which otherwise only are formed during metamorphosis to the sessile adult stage.
- Caenogeneses are evolutionary novelties which are adaptations of embryonic or larval stages, like the development of secondary gills in aquatic insect larvae, primarily absent in adults (nonterminal character addition).
- Reductions of larval characters or larval stages (nonterminal deletion): for example the lack of planktonic larvae in species of the Annelida, Mollusca or Crustacea which secondarily acquired direct development.
- Reduction of adult characters (terminal deletion): fetalization (retention of larval characters in adults), progenesis (speeding up of sexual maturity in larvae) and neoteny (sexual maturity of larvae due to retarded somatic maturation, see larval gills of the axolotl) can formally be interpreted as reduction of adult characters.

Other variations that could occur are terminal substitution of adult or larval characters, nonterminal substitution, reversals (Mabee 1993).

To know which developmental and physiological mechanisms cause these phenomena is not relevant for systematists in each case, they mainly have to be able to identify the plesiomorphic character state.

These deviations are largely considered with Nelson's rule (De Queiroz 1985).

Nelson's rule

This is a variation of the ontogenetic criterion, which does not require an exact order of the recapitulation: "… *given an ontogenetic character transformation, from a character observed to be more general to a character observed to be less general, the more general character is primitive and the less general advanced*" (Nelson 1978).

The meaning of "general" is ambiguous. The more general character could be the one occurring in the majority of species regardless of the ontogenetic stage (*commonness across species*), implying that "common is primitive". This interpretation is untenable because, for example, character frequency may be the result of rapid radiation, which is independent of the phylogenetic age of the

majority of characters of an organism (see also Watrous & Wheeler 1981). Another version would be that the most common character is the one most frequent across all ontogenetic stages (*commonness across ontogenetic stages*). In this case the interpretation would depend on the number of discernible stages, a parameter that has nothing to do with character evolution (see also Meier 1997). The more logical interpretation is that plesiomorphic states should be present in more ontogenies than novelties that were added later:

The more general character is the character state which occurs together with a second state in one individual (not necessarily simultaneously but during ontogeny) while the first character state is found without this second state in another individual. The second state then would be the phylogenetically younger specialization, the first one is the plesiomorphy. In contrast to Haeckel's biogenetic rule, it is not required that recapitulation follows a specific order. A more formal description of this argumentation is:

1) a homologous character ***M*** only occurs in the species or the group of species ***A*** in state $\boldsymbol{M_1}$,
2) while in the species or group of species ***B*** it occurs in the states $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$.
3) Consequently, the state $\boldsymbol{M_1}$ is the more common and more primitive one.

The ontogenetic appearance of characters in the order that corresponds to the historical evolutionary events *is not necessary* when Nelson's Rule is applied, however, the apomorphic state should have evolved historically by addition of new details (*terminal addition*; the criterion is not applicable when the new state is similar to the old one due to reversal or multiple substitution). The wider distribution (in the sense explained above) of the more primitive character follows inevitably from the structure of the phylogenetic tree (Fig. 114). If the primitive state is not preserved in single cases, or developmental stages are missing, the rule will not be applicable. However, a misinterpretation can be avoided. Fetalization, progenesis or neoteny result in the absence of adult characters. As Nelson's Rule does not allow statements on the course of ontogeny, but only serves the evaluation of characters, neoteny is not a source of errors: the circulatory system of the axolotl and the external gills, for example, are recognized as more primitive than the character states of the non-neotenic species of *Ambystoma*; the point of appearance during ontogeny and the physiological and genetic causes of the ontogenetic development or the reduction of gills are irrelevant. For phylogenetic analyses the correct statement is obtained that in neotenic salamanders the gills themselves are not autapomorphies, but the apomorphies are those mutations which prevent the normal reduction of gills during metamorphosis.

The application of the ontogenetic criterion for the determination of character state polarity requires that (1) the observed developing larval or embryological structure is homologous to a functioning adult or larval organ, and that (2) variations of homologous characters did not evolve from duplicated copies (homonomies), because otherwise when in one species a copy is lost, the other copy misleadingly appears to be a specialization (= a derived character state).

As for most characters the ontogenetic development is not known and additionally many molecular characters (especially genes) do not pass through an ontogenetic modification, in practice the ontogenetic criterion is of secondary importance for the determination of character state polarity.

The case of an analysis of recapitulation to prove the *presence* of characters which are important for the systematization of a taxon is different (examples for important apomorphies: the nauplius larva as constitutive character of crustaceans occurs in the otherwise strongly modified barnacles; the anlage of segmental paired coelomic sacs in Prot- and Euarthropoda as character of the Articulata; these sacs are not present in adult arthropods). The study of recapitulation is also important for the homologization of characters with the criterion of continuity (ch. 5.2.1). It yields evidence on the origin and evolution of characters.

5.3.6 The paleontological criterion

If a character of a monophyletic group occurs regularly in one state in older fossils and in another state in younger fossils or in recent species, the state in the older fossils is the plesiomorphic one (Hennig's geological character precedence; De Jong 1980).

This argument should not be accepted uncritically. The expectation that phylogenetically old or-

Fig. 115. Example for supporting positions in an 18S rDNA alignment. In this case the split between vertebrates (upper 4 rows) and invertebrates occurs in 105 alignment positions. Only 54 of these are shown. The upper row of numbers gives the sequence positions in the alignment used. Taking the vertebrates as ingroup, the nucleotides of the ingroup on the white background are synapomorphies (= ground pattern characters of vertebrates), and nucleotides on the white background seen in outgroup sequences are chance similarities shared with the ingroup ("noise").

ganisms have primitive characters is generally well founded, but there are two sources of errors that must be considered:

1) Fossils can show autapomorphies, with a corresponding plesiomorphic state occurring in younger taxa. For example, the small size of the wings of the primitive and toothed but flightless bird species *Hesperornis regalis* (Fig. 86) and *Baptornis advenus* is undoubtedly a secondary adaptation to a mode of life similar to the one of extant flightless cormorants and therefore an autapomorphy. However, the presence of teeth can be interpreted to be a plesiomorphy. Character analysis with outgroup comparison allows the correct evaluation: wings suitable for flight are already present in *Archaeopteryx*, teeth exist in other fossil birds and in other tetrapods. Additionally, it can be shown that "tooth genes" are also present in recent birds (Kollar & Fisher 1980), even though they are not expressed.

2) The stratigraphic (geochronological) sequence in which fossils are found does not always agree with the historical sequence of origin of taxa. For example, it must be assumed that the colonial Scleractinia (stony corals), which occur in the fossil record since the Middle Triassic, descended from solitary Anthozoa which did not possess a skeleton. The probability that the soft body of these ancestral species has been fossilized somewhere and that it will be discovered is very low (Veron 1995).

5.3.7 Phenomenological determination of character state polarity in nucleic acid sequences and asymmetry of split-supporting patterns

Sequences regarded as "patterns" containing traces of evolutionary events differ in no essential way from morphological characters and can be analysed with the same methodological approach-

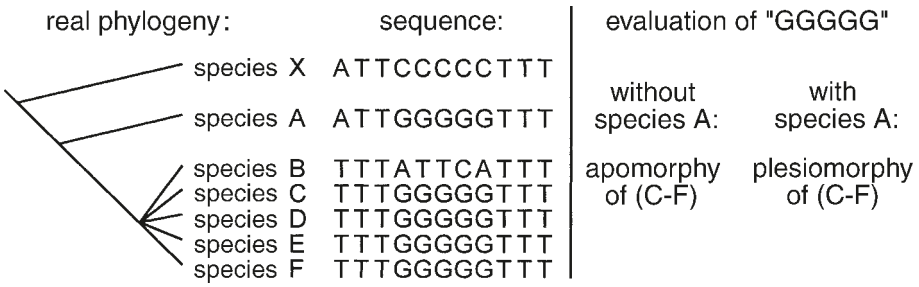


Fig. 116. The correct phylogenetic interpretation of character states requires knowledge of characters of many species. Without species **A,** monophyly of the species group **C-F** would be erroneously proposed. Note that species ***B*** has more substitutions than other species, it is a "long-branch-taxon"; several plesiomorphic states eroded due to subsequent substitutions.

es. An alignment is nothing else but a data matrix in which species and their characters are listed (Figs. 115, 116). If the data are phylogenetically informative a hierarchical order of species groups can be reconstructed from such a matrix. The character state polarity, however, has to be searched for deliberately. This search can be omitted when outgroups are determined using different methods of tree reconstruction. However, the sources of error implied with this "outgroup addition procedure" are the same as in phenetic cladistics (ch. 5.3.3, 6.1.11).

In order to determine character state polarity *before tree construction*, a character analysis has to be performed in a similar way as explained for morphological characters in ch. 5.3.2. In contrast to morphological characters, where a compact organ evolves *de novo* or may occur with modifications, one rarely finds characteristic longer sequence sections (signatures) that can be coded as apomorphies (Fig. 98). Usually the apomorphies are single nucleotides in substituted positions, which are scattered over the length of the sequence. The individual position is not very informative, but the sum of these positions gives us the patterns that have to be analysed (compare Fig. 115).

To find putative apomorphies in sequences without reference to a tree, the following steps are required:

- Search for sequence sections ("signatures") or for all single positions with nucleotides that occur in one or several, but not in all species. These positions all support the same bipartition of the set of species (Fig. 115). Name the species or group of species that share the same nucleotide per position the putative "ingroup". These individual positions and signatures are "supporting positions" for the putative ingroup.
- Name all other species the "outgroup".
- Check the sample of species: there should be enough outgroup sequences to guarantee that in relation to the putative ingroup closely related as well as distantly related species are present. It is justified to consider for this purpose well corroborated sections of the phylogenetic tree. Only those characters which have a different state in the ingroup than in the outgroup are used for the analysis.
- Compare the number of ingroup-supporting positions (the "signal") with the background noise that is caused by chance similarities, as described in ch. 6.5.1. Given the signal is markedly higher than the background noise, character states (nucleotides) which are characteristic for the ingroup are probably apomorphies.

The analysis gets complicated by the fact that deviations from the ground pattern can occur frequently in the ingroup and produce "noise" or "inconsistencies" in individual species. This is so because ancestral sequences continue to evolve after a speciation and ground pattern characters are substituted in single species. Since one wants to extract as much information as possible from the sequences, this noise has to be taken into account (more details in ch. 6.5.1).

In a partition separating a monophylum from a larger outgroup (as in Fig. 115) in most cases the ingroup taxon is historically younger than the

oldest outgroup lineage. This implies that more time is available for the occurrence of substitutions within the outgroup than within the ingroup. This effect can be seen in Fig. 115, where split-supporting positions are noisy in the outgroup, but not in the ingroup. For this reason we will find in many cases an *asymmetry of split-supporting patterns*.

In principle, patterns of supporting positions can be analysed without any previous knowledge about the structure of the tree. However, as in comparative morphology, aspects of the evolution of patterns can be derived more easily from the data (sequences) when it is known with certainty which of the species are not members of the ingroup. This usually is no obstacle in practice, because in each case some organisms are known which nobody would consider to be close relatives of the ingroup. Molecular systematists at the moment only have a limited number of sequences at their disposal. It is important that several taxa are contained in the outgroups, so that plesiomorphies of the ingroup are not misinterpreted as apomorphies (Fig. 116).

Amino acid sequences can also be analysed in a similar way.

6. Reconstruction of phylogeny: the phenomenological method

In general, tree constructing methods should have the following properties:

- efficiency (the method should be fast),
- power (a correct result should be obtained with a minimal number of data),
- consistency (increasing the dataset the result will converge on the correct tree),
- robustness (minor violations of the method's assumptions have no drastic effects),
- falsifiability (it should be possible to know in which cases the method is not applicable).

Methods differ in their properties: parsimony methods are for example slower than distance methods, but they often are more robust. Falsifiability based on statistics is a problem that has not been solved in a satisfactory way. Furthermore, since the correct tree is not known in most cases, power and consistency are not easily demonstrated. Therefore the plausibility test is recommended to check if the result fits to other data (ch. 10).

The following steps have to be distinguished for the reconstruction of phylogeny in a phenomenological approach:

- evaluation of characters (phenomenologically, or considering models, *a priori,* or *a posteriori*),
- construction of topologies (from combinations of all terminal taxa or from patterns present in the data),
- selection of the best supported branches of the optimal topology or of equally good topologies based on the preferred optimality criteria ("most parsimonious" topology, "most probable" topology) and test of the topology by comparison with other datasets and by examination of the plausibility of the evolutionary scenario implied by the phylogeny.

Since all three steps are linked to each other in different ways, depending on the method used, they will be explained in other chapters (where required by the method).

The comparison of homologous characters on principle allows the uncovering of a hierarchical order of taxa. If character polarity is unknown this order is primarily not polarized, the corresponding tree is unrooted. In which way this order can be disclosed will be explained in the following paragraphs (ch. 6.1.8). A basic principle necessary for each illustration of hierarchies of character states or of taxa is the identification of groups of taxa that share similarities. We assume that in a preceding step homologous and analogous similarities have been distinguished (ch. 5) and that only putative homologies are used for further work.

The selection of relevant groups represented in a dataset shall be illustrated with a short artificial alignment (Fig. 117):

position:	1	2	3	4	5	6	7
species A	A	A	T	A	A	A	A
species B	A	A	A	T	A	A	A
species C	T	A	A	A	T	A	A
species D	T	T	A	A	A	T	A
species E	T	T	A	A	A	A	T

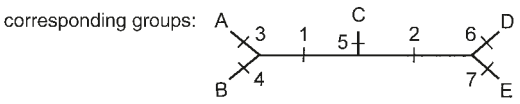


Fig. 117. Identities in alignment positions support the distinction of groups of taxa. The numbers in the tree correspond to alignment positions with character state changes (characters, details in frame homologies).

Each position of this alignment (Fig. 117) can justify the separation of a terminal species or a group of species from all others. For example, character *A* in position 1 has the same function for the support of group (A, B) as character *T* for the group (C, D, E).

Each character substantiates a group in the following way: if a character (or a state) is present in some species but not in others, the group with and the one without the character are delimited from each other by this fact. If a character is a "state" (a detail within a larger pattern) and exactly 2 states can be distinguished in the dataset, each state characterizes one group. Thus nothing must be known about character state polarity to find groups. The graph constructed from mutually compatible groups (see ch. 3.2) visualizing the corresponding encaptic order is an "unpolarized" or "unrooted" tree. Of course, in this case only part of the groups are putative monophyla.

6.1 Phenetic cladistics

The method of data analysis called **"cladistics"** in the English literature has mutated in the past years, there exist different methodological approaches. The methodology that does not include *a priori* character analysis will be called "phenetic cladistics" in the following, in order to stress the difference from phylogenetic systematics or, better, to "phylogenetic cladistics" (in German: "Phylogenetische Systematik", as founded by Willi Hennig, ch. 6.2), which can use the same algorithms.

Phenetics is the description and comparison of visible structures. It became common practice to use the word for methods which did not imply the independent search for homologies and character state polarities, such as the numerical description of allometries (allometry: variation in proportions of body parts). Phenetic methods of systematics were originally developed under the name "**numerical taxonomy**" (e.g., Sokal & Sneath 1963), to evaluate similarities quantitatively. These methods comprise the transformation of a species/character matrix with arbitrarily selected parameters into a similarity matrix containing statements on the similarity of the species. Hypotheses of homology are not tested, the process of character evolution is not considered. Cluster analyses allow the partial graphical representation (omitting conflicting evidence) of degrees of similarities in the form of dendrograms.

Numerical taxonomy, as originally conceived, is generally not applied any more. It has been replaced by more efficient methods. Simple distance methods are also phenetic methods in which the criterion for the calculation of similarities is, for example, the number of differences in DNA sequences. For example, a phenetic statement on morphological characters is that a barnacle at first sight is more similar to a limpet than to a crustacean. The correct phylogenetic statement would be that the habitus of barnacles shows convergences to limpets due to adaptations to the same environment, but they do not share unique apomorphies. Model-dependent distance methods do not belong to phenetics in the strict sense, because assumptions on processes of evolutionary character transformations (substitution rates) are used, which can be inferred from visible similarities (ch. 8.2).

Based on experiences gained during the search for numerical methods, cladistic algorithms were developed that do not evaluate some quantitative measure of similarity but consider only the presence or absence of discrete characters. The treatment of individual characters as units that measure similarity (without evaluation of character quality) is retained in phenetic cladistics. Many of the contemporary scientists applying this method call themselves "cladists".

Phenetic cladistics is also called **transformed cladistics**. This refers to the modification of Hennig's method and the introduction of his terminology. It is a methodological approach in which groups of species are characterized with putative apomorphies, without the necessity for a theory-based justification (such as an analysis of the probability of homology and of character state polarity) for the use of these characters (Nelson & Platnick 1981). Species are defined as classes showing specific characters. It has already been discussed that this species concept can hardly be useful for systematics (ch. 2.3: see species concept of Cracraft 1987). One aspect of the methodological approach, however, is epistemologically well founded and is also advocated in this book: when phylogeny is to be reconstructed objectively and independently of any *a priori* assumptions on the course of evolution, character patterns ("traces left by evolution") have to be analysed with methods which do not require *ad hoc* assumptions, if possible.

Essential axioms of phenetic cladistics are:

- Characters are homologous when they are congruent (**criterion of congruence**); congruence means in this case that the distribution of character states among taxa of the data matrix is compatible with a specific most parsimonious dendrogram. The homologous states are those that change only once on branches of the tree. (This implies an *a posteriori* determination of homology, see ch. 6.1.10, 6.1.11).
- The optimality criterion for the reconstruction of phylogenetic trees has to be the principle of parsimony (see ch. 1.4.3 and 6.1.2).
- The plesiomorphic character state of species placed within a monophylum is the one which occurs in the most closely related taxa of the outgroup (= cladistic determination of character state polarity by outgroup addition, see ch. 5.3.3).

The disadvantages of the cladistic determination of character polarity have been discussed in ch. 5.3.3, the problems of the *a posteriori* homologization are summarized in ch. 6.1.10.

For many cladists it follows from the criterion of congruence (see ch. 5.1.1) that an *a priori* weighting of characters is unnecessary or not possible (e.g., Goloboff 1993), a point of view which prevents the application of the hypothetico-deductive method and therefore has to be rejected (Bryant 1989).

In comparison to Hennig's original method, the **advantages** of phenetic cladistics that should be mentioned are:

- the possibility to use exact and fast computer programs,
- implying the chance to find optimal topologies for large datasets,
- the detection of all alternative optimal topologies, when homoplasies are present or when data do not contain enough information to support a single optimal tree.

These advantages can also be used in an updated Hennigian analysis, in *phylogenetic cladistics* (ch. 6.2), avoiding the drawbacks of the phenetic analyses. The criterion of congruence then serves only as a test for previously substantiated hypotheses of homology for apomorphies (ch. 1.4.3).

First suggestions on how to select most parsimonious topologies came from Edwards & Cavalli-Sforza (1963), who proposed the "**minimum evolution method**": the topology which requires the smallest "amount of evolution" has to be preferred (ch. 8.2.7, 14.3.8). The authors referred to the probability that rare characters originate only once, assuming that this reflects the "parsimony of evolution". This concept is not the same as "maximum parsimony" in MP-methods, where the principle of parsimony is a *methodological* principle (see Edwards 1996 and ch. 1.4.5). In contrast to the MP method, the "minimum-evolution"-method relies on distance estimates that are calculated for pairs of terminal taxa to estimate "branch lengths". Discrete characters are not compared. It is assumed that the tree with the smallest sum of branch length estimates is most likely to be he true one. First suggestions for algorithms were published by Kluge & Farris (1969) and Farris (1970), who explicitly referred to W. Hennig. A compa-

linear transformation series of a character:

species:	A		B		C		D
character state:	M_a	←→	M_b	←→	M_c	←→	M_d
coding:	1	←→	2	←→	3	←→	4

matrix:

species/ character	binary coding: M_a	M_b	M_c	M_d	transformation coding M
A	1	0	0	0	1
B	0	1	0	0	2
C	0	0	1	0	3
D	0	0	0	1	4

Fig. 118. Coding of a transformation series. The transformation coding requires the assumption that the different characters (or character states) evolved step by step from each other. However, the depicted binary coding of the presence of characters does not, it would not describe character evolution. A binary coding of the transformation series is shown in Fig. 119.

rable algorithm had never been suggested by W. Hennig. The first methods were not very popular because the calculation of branch lengths for all alternative topologies is very time-consuming and therefore a fast distance method was preferred, the "neighbour-joining"-algorithm (see ch. 8.2.7). In the meantime more efficient algorithms have been developed which require little computing time (Gascuel et al. 2001).

6.1.1 Character coding

The basis for a cladistic analysis is the data matrix (Fig. 118, 119). To compile the matrix one has to define the character, its states, and one has to assign states to terminal taxa. The delimitation of characters and states can be called the *coding s. str.*, the assignment of states to taxa then is the character *scoring* (Jenner 2002). Since both steps are usually carried out simultaneously, in the following the term "coding" is used in its wider sense.

matrix:

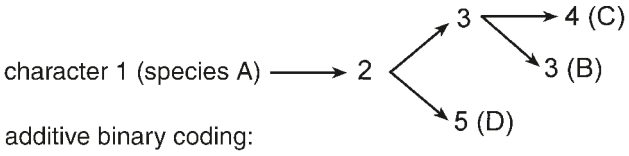
species/ character state M_i	M_a	M_b	M_c	M_d
A	1	0	0	0
B	1	1	0	0
C	1	1	1	0
D	1	1	1	1

Fig. 119. Additive binary coding for the linear character evolution shown in Fig. 118.

The data matrix contains entries representing character states for each terminal taxon. If the character states of a single *taxon are written in rows*, then the various *states of a character* can be found in a *column* (Figs. 117, 118). Note: in this way of coding, a character consists of a frame homology, represented by the column itself, and the detail homology is the character state represented by entries in the column. Here the homology of the frame is presupposed and usually not tested in the course of further analyses (see also ch. 6.1.10). Homology of the "frame" must be assessed in an independent previous analysis. The frame homology corresponds to the positional homology in alignments of sequences (ch. 5.2.2.1).

For each *variable* character at least two statements are possible ("present" or "absent", equivalent to "state 1" or "state 2" of the character), which are represented with symbols (usually "1" and "0"). This coding implies no polarity! When transformation series are known (s. Fig. 77), two equivalent possibilities to code these can be used: a succession can be represented with a series of numbers, whereby it is assumed that, for example, character state "3" previously passed through states "1" and "2" (transformation coding in Fig. 118). Alternatively, the same fact can be coded in a binary mode (Fig. 119).

The interpretation of **binary coding** of transformation series as in Fig. 118 (left columns) re-



additive binary coding:

species/ character	1	2	3	4	5
A	1	0	0	0	0
B	1	1	1	0	0
C	1	1	1	1	0
D	1	1	0	0	1

Fig. 120. Branched transformation series of a character.

quires an instruction explaining which steps are allowed (character transformations: M_a/M_b is allowed, but not M_a/M_c). This instruction is redundant with additive binary coding (Fig. 119).

This **additive coding** implies that character M_a also occurs in species B and that the novelty M_b has been added (Fig. 119). In the case of M_b being a younger variation of M_a the coding would be the same. In case species A is the phylogenetically older one this coding would mean that for species B the character M_a is a plesiomorphy and character M_b is an apomorphy of B + C + D.

The divergent evolution of a character in sister taxa can be illustrated with **branched transformation series** as in Fig. 120.

It is recommended to describe branched transformation series with additive binary coding (Fig. 119, 120), because in this case further instructions for the interpretation of the series of numbers can be omitted.

Whenever a transformation series is known we are dealing with "**ordered characters**". "**Unordered characters**" are present when it is not known in which chronological order the states evolved from each other or whenever in principle transformations are expected to occur between any alternative states. Ordered series can be described without establishment of character state polarity. The series of mouthparts in Fig. 77, for example, could be read in both directions if the plesiomorphic state is not known. Unordered characters are for example the nucleotides of a DNA sequence which are classified as states of the character "sequence position": each nucleotide can be substituted by any other one (Fig. 121).

If the succession of character states is known, it is convenient to mark them as "ordered characters" in the computer programs selected for cladistic analyses. With the specification of the chronological succession additional information is gained for phylogenetic analyses. This information should not be ignored light-heartedly.

Additional information can be gained with the determination of the **order of character states** or, in short, **character state polarity** (see ch. 5.3). If polarity is known for all characters used for a cladistic analysis, the **Dollo** algorithm is chosen (6.1.2.3), whereas the **Wagner** algorithm does not require a polarization. Character series with known polarity are called **polarized**, without polarity they are called **unpolarized**. (In princi-

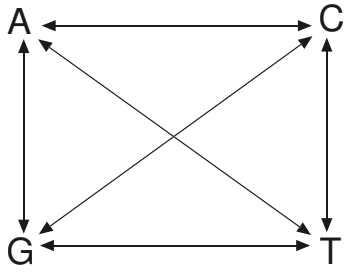


Fig. 121. Possible transformations of an *unordered* character exemplified with four nucleotides occurring in a position of a sequence alignment. In comparison to morphological characters, the position is the character (or the frame homology), the nucleotide is the state (the detail homology), whereby each state can evolve from any other state.

species/ character	1	2	3	4	5	6	7	8
0	-	-	-	-	-	-	-	-
A	+	-	-	-	-	-	-	-
B	+	+	+	+	+	-	-	-
C	+	+	+	+	+	-	-	-
D	+	+	+	-	-	+	+	+
E	+	+	+	-	-	+	+	+

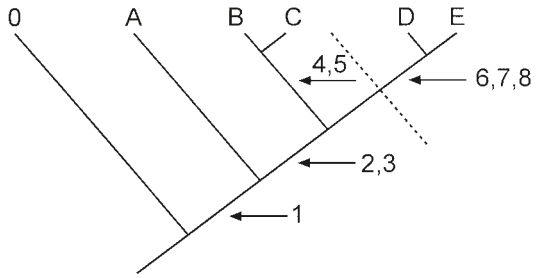


Fig. 122. When characters have a mutually compatible distribution, the character matrix contains the same groups of species as the corresponding most parsimonious dendrogram. In this data matrix "–" means that a character is missing, "**+**" indicates the presence of a character. In order to establish the correct root of the dendrogram, all states coded with "–" have to be the plesiomorphic ones, otherwise the topology is unrooted. An ideal data matrix has no homoplasies.

ple, trees with fixed character polarities can also be called polarized trees, they are in fact rooted by determination of the direction of character evolution.)

States of an ordered, polarized character:

$$1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4$$

States of an ordered, unpolarized character:

$$1 \Leftrightarrow 2 \Leftrightarrow 3 \Leftrightarrow 4$$

Missing characters: when detail homologies are not described for a terminal taxon, this lack of knowledge is usually indicated in a data matrix with a question mark "?". The question mark can represent each of the character states which are found in other taxa or some unknown new state. MP-algorithms automatically replace the question mark with that character state which contributes less to the increase of tree length. The topology of those tree sections which for some characters of terminal taxa have only question marks is determined by other characters that have defined states.

Inapplicable characters are often indicated with a dash ("–") but treated in the same way as missing data. Some snails, for example, have a flattened shell, others a long coiled shell, while slugs possess no shell. If it would be known that the slugs are derived from species with flattened shells, the coiled shell being the plesiomorphic state, three character states could be coded ("long coiled", "flattened", "reduced"). However, if the shell state of the last common ancestor of slugs is unknown, one would distinguish two different characters: "shell present or absent" and "shell shape", the latter being inapplicable for the slugs. In MP analyses a shell shape of the slug ancestor would be chosen that minimizes the tree length determined by the other characters of the matrix.

Whenever it can be shown that a detail homology is lacking secondarily, one should code this hypothesis with a third character state: 1→2 (evolution of a novelty), →3 (secondary reduction of the novelty).

Absence coded as a character state can be a problem, because absence has no structural complexity and therefore there exists no evidence for the homology of this state. This is true for morphological characters as well as for sequence data, where gaps are the equivalent. Often characters have to be coded as absence/presence characters (e.g., two pairs of thoracic wings present in Ptery-

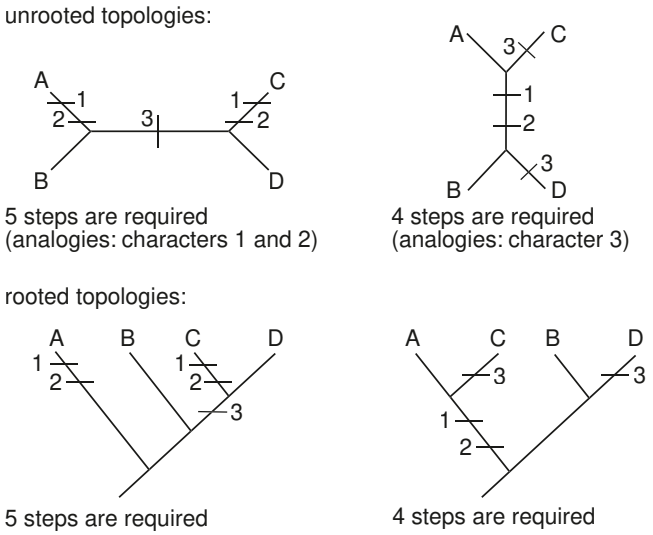


Fig. 123. Alternative topologies for the species A-D and the characters 1-3. The lines next to the numbers mean that a character changes its state or evolves *de novo* on the corresponding branch, without requiring a statement on character state polarity. Even if the root is not known, the selection of a most parsimonious topology is possible. In this example all characters are equivalent (equally weighted). When a topology has been chosen that is considered to be the most probable, the new character states on the inner edges (= novelties on potential stem lineages) are potential homologies.

gota, absent in all other organisms). To avoid that primary absence is used as a homology, the character has to be defined as a Dollo character (ch. 6.1.2.3). Gain of the novel character has to be weighted according to its relative complexity, while the loss should get a low or no weight.

Note: The coding of transformation series, polarity, and weighting of individual characters for parsimony analysis with the program PAUP (Swafford 1990) can be easily prepared with the program MacClade (Maddison & Maddison 1992).

6.1.2 The MP-method for tree construction

The methods explained in the following paragraphs are implemented in several computer programs, their application requires a conscious use and evaluation of the available data. It is particularly essential to be aware of the axiomatic assumptions which are implicitly required by these methods and one needs an estimation of the quality of the data, because the programs will always calculate for you some tree, even if you invent the data or using some chance numbers, or, which is frequently the case, when the data have a low information content.

The method of phenetic cladistics is limited to the "cladistic step" of phylogenetic tree construction (Fig. 123, see also Fig. 138). It has been claimed that it considers the *principle of parsimony* (ch. 1.4.5) objectively in a way that can be implemented in computer programs. Therefore, in the English literature the term "**maximum parsimony**" (**MP**-) method is now a synonym for the phenetic method of cladistic tree construction. Making use of these tools it is usually overlooked that in phylogenetics the principle of parsimony also has to be applied to character analysis: the most parsimonious explanation for the occurrence of identical complex characters in different organisms is that these patterns are homologous (compare ch. 5.1).

For the cladistic construction of trees the following laws have to be considered: when a dataset does not contain conflicts, i.e. when homoplasies are absent, then the structure of the most parsimonious dendrogram is unequivocally determined by this dataset (Fig. 122; for the term "homoplasy" see Fig. 78). In other words, the tree

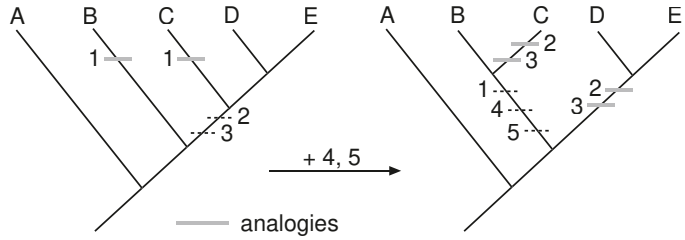


Fig. 124. Assume that the novelties 2 and 3 for the group C-E (left dendrogram) are characters, which are considered to be important apomorphies on the "true" tree. Such characters can forfeit their effect in MP-methods when convergent characters or chance similarities (characters 4 and 5 in the right dendrogram) are added to the original data matrix. Because the convergences outnumber the true apomorphies, the convergence of the original dataset (1 in the left dendrogram) and the added characters become synapomorphies (1, 4 and 5 in the right dendrogram), the true synapomorphies are degraded to analogies (2 and 3 in the right dendrogram). Even though the right dendrogram is the most parsimonious for the complete dataset, it does not represent the correct phylogeny.

is already contained in the data. However, the dendrogram does only represent aspects of phylogeny under the condition that the characters in the dataset are real homologies and when their polarity is known. Furthermore, the dendrogram is already a completely reconstructed image of phylogeny when for each monophylum just a single real apomorphy is known. (The problem is that we never can be sure which characters are real apomorphies! Therefore we have to operate with probabilities.)

When incompatible characters are present (see homoplasies in Fig. 78; Figs. 123, 193, 195), there may exist several most parsimonious dichotomous dendrograms which describe the structure of the dataset. The MP-method serves to find the "best" topologies, but cannot help to decide which of the alternative "optimal" topologies is the correct one.

The method consists of a search for the tree topology which requires the least amount of character changes. Each character state change (including the evolution of a new character or the reduction of a character) is counted as one step, the number of steps for the whole topology yields the "tree length". Also multiple changes of the same character are regarded to be equivalent and counted individually. Hereby polarity is not important, it is sufficient to record for a given branch of a topology which frame homology is modified or is added or reduced (Fig. 122). The "shortest tree" is chosen as a basis for statements on phylogeny. Phrased differently, each "**step**" causes "**costs**", the "cheapest" tree is favoured (in Fig. 123: 4 steps are "better" than 5 steps).

How to find the "cheapest" or shortest tree is explained in the appendix (ch. 14.2). A direct analytical calculation of the shortest topology is not possible. Instead, alternative topologies are constructed (if computationally possible, all combinations of the available taxa are considered, see ch. 14.2) to compare their lengths. "Calculation" means in this context the heuristic or exact search for the shortest tree, usually conducted with a computer program.

The cladistic MP-method makes it possible to distinguish topologies, even when the polarity of characters is not known (Fig. 123). The **polarization** or **rooting** is performed using polarized characters (ch. 5.3.2) or by cladistic outgroup addition (ch. 5.3.3).

The MP-method is very sensitive to variations of the number and weight of potential apomorphies (Fig. 124). For this reason it is important that the characters are weighted according to their probability of homology. If you prefer not to weight the characters, a few insignificant characters can drastically change the topology of a dendrogram. Many cladists, however, reject evaluating characters *a priori*.

This method implies that, in contrast to distance methods, **trivial characters** (autapomorphies of terminal taxa which do not show convergences) and plesiomorphies, which only occur in one taxon of the outgroup, have no influence on the topology, but nevertheless increase the tree length

species	MP-informative characters	trivial characters	constant characters
A	1 1	1 0 1	1 0
B	1 0	0 1 1	1 0
C	0 0	0 0 1	1 0
D	0 1	0 0 0	1 0

Fig. 125. Classification of characters according to their effect in MP-methods. Trivial characters are usually potential autapomorphies of individual species or of terminal taxa. In this example the MP-informative characters produce a split {(*A*,*B*),(*C*,*D*)} and the split {(*A*,*D*),(*B*,*C*)}.

when they are included in the count. All characters occurring with two or more states and with each state in more than one terminal taxon are effective. Such characters are called **parsimony-informative characters**. They produce in a dataset a split which separates groups of more than one species each. It is not required that the splits are compatible with each other to allow an analysis of the data.

Conditions for the use of the MP-method to reconstruct phylogenetic trees

- There must be good reasons to assume that all characters are **homologies**.
- All characters must have either **the same** estimated **probability of homology** or characters with high probabilities of homology must get a high weight.
- All characters have to be derived from a reconstruction of the **ground patterns of terminal taxa** or they must represent constant characters of terminal species.
- Terminal taxa must be **monophyletic**.

The conditions for the application of the maximum-parsimony method listed above have the function of **axioms**, which cannot be tested with the same **deductive method** (which is the search for the "most parsimonious tree"; for the role of axioms see ch. 1.4.2)! Violations against these axioms produce topologies which erroneously inspire confidence, because the usual tests combined with cladistic methods (ch. 6.1.9) logically cannot detect a source of error in the area of axioms.

Problems and sources of errors of these numerical analyses are presented in ch. 6.1.11 and ch. 9. The frequently debated statement that hypotheses of homology should be substantiated prior to the phylogenetic analysis will be discussed in ch. 6.1.10.

6.1.2.1 Wagner parsimony

The Wagner parsimony (= optimality criterion of Wagner 1961) is a method of cladistics to determine the length of a tree presupposing the following assumptions:

- all characters are reversible, meaning that the probability of the transformation of character states is independent of the direction of the transformation ($0 \rightarrow 1 = 1 \rightarrow 0$).
- Character states are ordered: a very derived state can only be reached stepwise through all intermediate states ($0 \rightarrow 2 = 2$ steps; see Fig. 118).

Assuming reversibility means that it is for example possible that in the course of evolution a compound eye could be reduced in one step to evolve *de novo* in the next "equally long" step. With this assumption the fact is overlooked that a destructive mutation is an uncomplicated event, while the evolutionary construction of a compound eye requires a large number of very specific mutations and millions of years of selection. Do not confuse this with a silencing of genes in some species and a reactivation of these genes later during phylogeny, which is a more probable assumption of reversibility.

Wagner parsimony is typical for the original approach of phenetic cladistics. Based on concepts of Wagner (1961, 1963), the cladists Kluge & Farris (1969) and Farris (1970) concentrated on writing computer programs which were suitable to calculate a dendrogram from a species/character matrix. This approach had such a great fascination for many scientists that the relevance of the complexity of characters as source of information was overlooked. The efforts of researchers were dedicated to improve the methods of tree reconstruction, but not to understand the evaluation of the quality of data.

The optimality criterion of the Wagner parsimony is "tree length". Starting with a data matrix with binary or additive binary coding of characters, tree length is calculated as follows:

- We start with a given topology whose length has to be determined.
- Determine for each character the state in inner nodes (this corresponds to the cladistic determination of character states in ground patterns, see Figs. 127, 131).
- Select the first character in the data matrix and count how many times the character state changes along the topology of the tree. Each change on an edge (branch) is a step (compare Fig. 123). The succession or direction in which the edges of the topology are analysed is irrelevant.
- Repeat this count for each character of the data matrix.
- Add all steps found, the sum is the "tree length".

As the polarity of characters is not considered, it is irrelevant whether the tree is polarized ("rooted") or not.

"Step": character state change on an edge of a given rooted or unrooted topology, independent of the polarity. If a character is weighted, the number of steps is the weight for a single character state change.

Tree length: sum of character state changes (= steps) of all characters of a dataset in a given topology

"Shortest tree": topology with the least number of character state changes

6.1.2.2 Fitch parsimony

The Fitch parsimony (Fitch 1971) is a variation of the above-mentioned Wagner parsimony. The fundamental assumption that character states are reversible is maintained. Additionally, there is the assumption that each character state can be transformed without intermediate steps into any other one and each transformation is equivalent. This corresponds to the use of **unordered characters** (number of steps for $0 \rightarrow 1 = 0 \rightarrow 2 = 1 \rightarrow 2$; Fig. 121). With this algorithm it is possible to analyse a DNA or protein sequence in which each character transformation is allowed (e.g., a single step for $A \rightarrow T$, $A \rightarrow C$, or $A \rightarrow G$).

6.1.2.3 Dollo parsimony

Dollo's law (ch. 2.7.1) is applied in MP-methods when the probability of loss mutations ("destructive reversals") is greater than the probability of the evolution of new characters (these could also be termed "constructive reversals"). Each complex derived character should therefore originate only once, each homoplasy would have to be explained with loss mutations (mutations re-establishing a previous character state) and thus would have the appearance of a plesiomorphy in comparison with the other corresponding character state; the putative plesiomorphy, however, originated secondarily. This procedure requires an *a priori* determination of the polarity of characters, or at least an estimation of the probability that loss mutations or the new evolutionary construction of characters took place. The characters of species C and D in the example (Fig. 126)

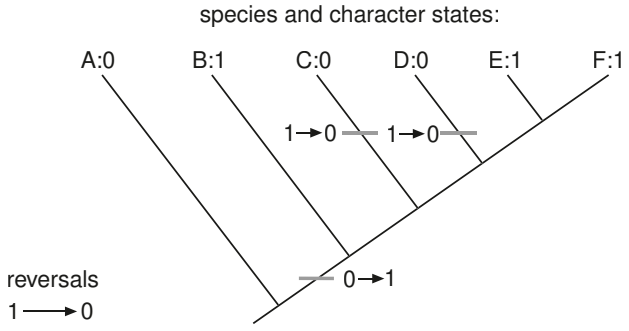


Fig. 126. Example for a character distribution in a given topology to explain Dollo parsimony, which allows the occurrence of destructive reversals (loss mutations). Destructive reversals produce apparent plesiomorphies, although each event (in the example above) is an evolutionary novelty.

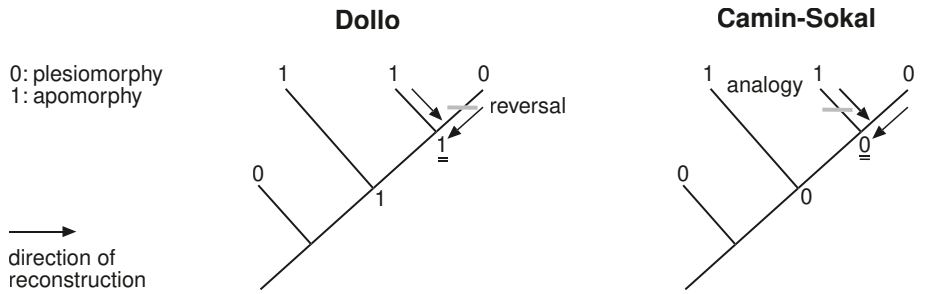


Fig. 127. Reconstruction of ground patterns with the cladistic MP-method, alternatively with the Dollo or the Camin-Sokal algorithm. The arrows show the direction of the reconstruction. The total number of character changes is the same in both cases, the selection of the algorithm for the determination of the most parsimonious topology is therefore not relevant. It can be seen that under Dollo conditions the path from one derived character of a terminal taxon (here coded with "1") along the edges of the topology to the next one does not lead through nodes (ground patterns) with plesiomorphies (coded with "0"). Even though tree length is the same, single branch lengths differ in both topologies. This might cause differences when divergence times are estimated using branch length information.

would be convergences which originated from loss mutations. In a simplified application of computer analyses without evaluation of characters (not recommended for morphological characters), Dollo parsimony means that "destructive reversals" to a former plesiomorphic state are allowed, but not "constructive reversals" that produce apomorphies. If single characters appear as presence/primary absence characters, it is wise to code them in such a way that absence cannot have the effect of a group-supporting character (an apomorphic homology). The step from absence to presence should get a high weight (depending on the character's relative complexity), and the reverse should have a very low or no weight (**Dollo character**).

A further variant is the **Camin-Sokal parsimony** (Camin & Sokal 1965)**,** which requires the assumption that the evolution of each character is principally irreversible. Reversals of any kind are excluded in this case. Each homoplasy then has to be explained with analogies or convergences.

The determination of character states in inner nodes differs in both methods (Fig. 127), because with Dollo parsimony it is assumed that a reversal took place, while with Camin-Sokal parsimony a plesiomorphy in inner nodes is conserved and the occurrence of analogies is postulated.

An unfounded generalization is based on the fact that the same algorithm is used for the analysis of the complete data matrix. Using computer programs it is recommended to code characters which probably are irreversible (e.g., the reduction of a compound eye) as such. A distinction of backward mutations that are not very probable (e.g., the evolution of new compound eyes in descendants of blind deep-sea crabs) and others which could occur more often (loss of pigment patterns, modification of bristle number, etc.) corresponds much better to the natural events, and is usually considered by systematists who use a phenomenological character analysis (ch. 5).

6.1.2.4 Generalized parsimony

The methods presented above are only suitable for special cases because they require special assumptions for the whole set of data, like equal weighting or the general reversibility of substitutions. Algorithms that allow **differential weighting** of character transformations for the reconstruction of dendrograms with the principle of parsimony are called "generalized parsimony methods". The consideration of differential weighting is possible with a cost matrix for each character. If in the character transformation series $1 \rightarrow 2 \rightarrow 3$, the evolution of each state from the preceding is weighted 8 times higher than the reversal, the following matrix results for all possible transformations:

old state	new state		
	1	2	3
1	-	8	2x8
2	1	-	8
3	2	1	-

Fig. 128. Example for a table with differential weights for character transformations (see text).

In this example the convergent occurrence of the transformation 2→3 would increase tree length by 8 steps, while the occurrence of a second reduction 3→2 would cost only 1 step. This is a way to code Dollo parsimony for this character. Should a reversal be impossible then it is weighted with the value ∞. With this method it is also possible to weight transversions and transitions occurring in the evolution of DNA sequences with different probabilities.

Disadvantages of these methods are the additional work necessary for character coding and the longer computation time. In the opinion of many cladists a further disadvantage is the subjectivity of weighting. However, in response to this it has to be pointed out that the assumption that all characters evolve according to the axioms of the Wagner or Dollo parsimony is also subjective, and, additionally, in most cases they are unrealistic. (Further explanations on the justification of weighting in ch. 6.1.3).

6.1.2.5 Nucleic acids and amino acid sequences

Sequences are usually analysed so that each position of the sequence is counted as an individual character (more details in ch. 6.3). A further possibility is the use of specific insertions, deletions, or series of substitutions as single complex characters, provided such characters can be found.

In the framework of phenomenological phylogeny inference, the analysis of DNA sequences implies (as in the case of the analysis of discrete morphological characters) that those characters which are implicitly or explicitly used to substantiate monophyly of a group of species should be apomorphies. The probability that the latter are not chance similarities can be estimated with an *a priori* analysis in a way similar to the evaluation of morphological characters (with spectra: see ch. 6.5). For an *a posteriori* analysis (after tree construction) it is recommended to print out and to evaluate those positions of an alignment which contain potential apomorphies for the monophyla of a well chosen dendrogram. In an ideal case these positions would have two character states: the plesiomorphic state in the outgroup, the apomorphic one in the ingroup (Fig. 115). The more deviations per position occur (absence of the apomorphy in single taxa of the ingroup, convergence to this apomorphy in single taxa of the outgroup), the more variable is this position and the lower is the probability that a real apomorphy can be identified.

Comparing genes coding for proteins, there are several possibilities to make use of the cladistic criterion of parsimony:

– Comparison of the coding regions of DNA sequences: each base substitution can be considered for phylogenetic analyses, including also synonymous substitutions which are not exposed to selection pressure. Codon positions can be weighted differentially according to their variability (compare ch. 2.7.2.4 and Fig. 48). A simple method consists of the elimination of the third or the third and second codon position from sequences by way of trial, to test the effect of individual positions on the quality of bootstrap values or on signal-noise ratios, for example.

When only the amino acid sequences are available:

– Comparison of amino acid sequences: only the base substitutions which are not synonymous (synonymous substitutions: see Fig. 48) become effective when the proteins are analysed. Consequently, less information is available than for the DNA analysis (for 3 nucleotides only one amino acid). But also in many cases multiple substitutions (e.g., in third codon positions) are omitted. This can be of advantage to eliminate "noise". In the simplest case only the differences of amino acid sequences are evaluated, without considering the minimum number of substitutions necessary at DNA level to transform the codon of an amino acid into another one (Eck & Dayhoff 1966).
– Count of the number of base substitutions which are necessary in order to convert the

	character (*weight)	
species A	1 (*5)	0 (*1)
species B	1 (*5)	0 (*1)
species C	0 (*1)	1 (*1)
species D	0 (*1)	1 (*1)

is the same as:

	character (*weight)					
species A	1 (*1)	1 (*1)	1 (*1)	1 (*1)	1 (*1)	0 (*1)
species B	1 (*1)	1 (*1)	1 (*1)	1 (*1)	1 (*1)	0 (*1)
species C	0 (*1)	0 (*1)	0 (*1)	0 (*1)	0 (*1)	1 (*1)
species D	0 (*1)	0 (*1)	0 (*1)	0 (*1)	0 (*1)	1 (*1)

Fig. 129. The weight of homologies of the species A+B that is effective in MP-methods is the same in both tables, although the table at the bottom contains more characters. In the lower table character 1 of the upper table was entered five times, which has the same effect as giving the transformation 0→1 in the upper table the weight five. The same effect can be obtained when five detail homologies of the frame character number one that support the split {(*A*,*B*)/(*C*,*D*)} are found and coded separately. The latter procedure is better justified than *ad hoc* weighting.

codon of one amino acid into another (compare PHYLIP program, Felsenstein 1993).

– Weighting of amino acid substitutions according to the chemical properties under the assumption that a change of the chemical properties is less probable, because it is subject to stronger selection pressure. Empirical data can be used to find specific weights for pairs of amino acids (compare Dayhoff et al. 1978, ch. 5.2.2.10).

Bear in mind that weighting according to the minimum number of substitutions required for a codon transformation is a pattern analysis, i.e. an evaluation of the visible differences, whereas weighting according to chemical properties of amino acids implies assumptions on selection processes. When weighting is based on the number of changes of a character along a given topology and then a new topology is estimated using the reweighted characters, the method is circular and therefore not admissible (see ch. 6.1.4).

On principle, it is also possible to use the MP-method in combination with models of sequence evolution. The number of potential apomorphies can be corrected with the number of non-visible multiple substitutions when Hadamard-conjugation is used in order to estimate generalized distances (instead of pairwise distances) between groups of taxa (Charleston et al. 1994, see also ch. 14.7).

6.1.3 Weighting and the MP-method

When characters are weighted as described in the following, the consequence is that in reality a weight is assigned to the *character state change* or to the appearance of evolutionary novelties in the 'true' and rooted tree. Therefore, the probability of homology should always be estimated only for those *detail homologies* which change or appear *de novo*, but not for a complete *frame homology*. To weigh the transformation of an incisor into a long tusk it is not correct to weigh the homology of the complete jaw or of the relative position of the incisor; the only relevant question in this case is whether the elongation of the incisor is a homology in two or more species. Weighting can be justified with the phenomenological approach (probability of the correct identification of homologies) or with the modelling approach (probability of the evolutionary origin of identities).

The topology of a dendrogram obtained with the cladistic method depends solely on the ratio of the number of parsimony-informative characters supporting different groupings and is primarily independent of the quality of characters. When characters are **weighted** according to their estimated quality (probability of homology), a character of higher weight has the same effect as when increasing the number of characters considered for a taxon (Fig. 129).

The phenomenologically working systematist can introduce an estimated relative ratio of probabilities of homology into the analysis by weighting a "valuable" character higher or entering it several times in the data matrix. An objective method consists of the identification of every discernible detail of a homology as exactly as possible, and to use each identified *homologous detail* as a character or to weigh the larger homology with the number of shared details.

The advocates of phenetic cladistics explicitly exclude a phenomenological character analysis and also *a priori* weighting, because ignoring the law discussed in ch. 5.1.1 (see criterion of complexity) it is maintained that statements on homologies gained from character analyses are unfounded *ad-hoc* hypotheses. Patterson's claim (1988) "the … most decisive test of homology is by **congruence** with other homologies" means that a hypothesis of homology has to be justified with the "fit" of a character (or state) to a "most parsimonious" dendrogram: it should emerge on a single inner edge and is a homology for the monophylum separated by this edge. Therefore, first of all a dendrogram has to be selected and only afterwards statements on the homology of characters are possible (Fig. 139). The reference to a dendrogram implies that for each character a polarity is also determined instantly when the topology is rooted, without consideration of character transformation probabilities. The problems of the cladistic homologization are discussed in ch. 6.1.10.

Attention: when character transformation series which develop stepwise from taxon to taxon are used, each state change of this series should be evaluated separately. It is a mistake to weigh the character globally, because this implies that all character state changes have the same probability of homology. Some computer programs allow weighting of each character state change separately (compare Fig. 128). – Also note that for **unordered characters** weighting may have a different effect than for ordered characters: an ordered character would add for a change from state 0 to state 2 two steps (0⇒1⇒2) multiplied by the weight, while an unordered character yields only one step (0⇒2).

6.1.4 Iterative weighting

Farris (1969) suggested a successive weighting procedure (also called *successive approximations weighting*) consisting of the following steps:

- construction of a dendrogram with the MP-method and with equally weighted characters,
- selection of values for weighting on the basis of character distribution in the most parsimonious tree: those characters which are distributed like synapomorphies get a higher weight, weight of homoplasies is reduced. The consistency index or retention index of the character can be used as a basis for weighting (ch. 6.1.9.1).
- Construction of a new MP-dendrogram with the new weights.
- Repetition of the weighting procedure now based on the new dendrogram.
- Repetition of the whole procedure until the topology of the dendrogram does not change any more.

The result of this method is very sensitive to the structure of the initial dendrogram and depends of the weights in the first data matrix, because during successive weighting especially those characters that are fitted to the topology of the first dendrogram are strongly "rewarded". Therefore iterative weighting is not a convenient method to find hypotheses on homologies, it is a **circular method** when used to identify homologies (*contra* Wenzel 2002)**.**

The circularity is easily demonstrated: trees should always be constructed from homologous characters and from character states that have a high probability of homology. In cladistics, differences of probability of homology are expressed by differential weighting. However, using successive weighting, initially, characters have the same weight and then weight is adapted depending on character fit to a tree, increasing the weight with decreasing degree of homoplasy. The resulting most parsimonious tree will be used to discern between non-homologous and homologous characters! So, weighting and identification of homologies is topology dependent and not based on independent evidence for character quality (see also ch. 6.1.10).

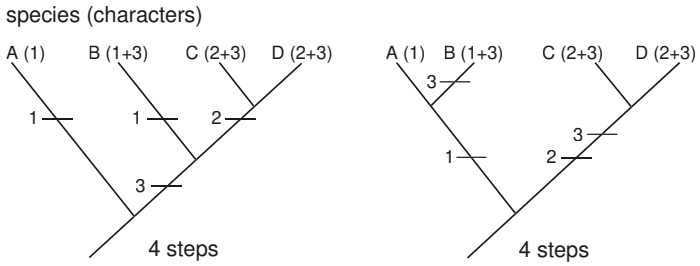


Fig. 130. In this example, there are two equivalent most parsimonious topologies for the same character distribution. Either character 1 (first topology) or character 3 (second topology) are homoplasies. It is obvious that homoplasies are contradictions in hypotheses of phylogeny (s. also Fig. 78).

The same problem occurs in methods which do not need iterations but provide weights of characters according to the number of homoplasies of individual characters, as in *implied weighting* (see e.g., Goloboff 1993). One can use different topology-dependent statistics (e.g., consistency index, number of character state changes) for weighting in the framework of maximum parsimony (e.g., with PAUP; Swafford 1990). Weights can, for example, be calculated with MacClade (Maddison & Maddison 1992). Implied weighting is fast because weights are defined during the first tree search. However, these methods are not reliable, because the weights depend of the topology and do not take into account the quality of individual characters.

6.1.5 Homoplasy

Homoplasies are characters whose distribution among taxa is not compatible with a dendrogram (Fig. 78, 130, 193, 195). As long as one has not decided which dendrogram is the most parsimonious one, or which topology is favoured by the systematist, the incompatible character is neither to be called a homology nor an analogy. This is why a separate term is used for this fact. It is a mistake to synonymize homoplasy with "analogy" or "convergence", as done by many cladists.

When a character has *n* states, it does not form homoplasies in a given dendrogram if this shows all character state changes on branches and if a total of *n-1* state changes occur in it. Whenever the number of state changes is larger, then homoplasies are present. Often one can find two or more equally parsimonious topologies due to the presence of homoplasies (Fig. 130).

It is obvious that homoplasies cause contradictions in phylogenetic hypotheses and therefore weaken or falsify hypotheses. There are different possibilities to consider homoplasies in further analyses (see also Siebert 1992):

- by analysis of the probability of homology (ch. 5.1) in order to assign new character names to different characters that were discovered to be analogies, and to exclude uncertain characters from the dataset, or to reweigh characters.
- by visualization of the presence of conflicts, for example with a consensus dendrogram (see ch. 3.3) or, even better, with a network diagram (s. ch. 3.2.2, ch. 14.4, Figs. 55, 56, 195) or a spectrum with incompatible splits (Figs. 153, 154).

When homoplasies occur in a dendrogram that probably depicts phylogeny correctly, the following causes should be considered (Givnish & Sytsma 1997):

- evolutionary convergence (similarity due to adaptation to the same environmental factors),
- analogy (chance similarity or "recurrence"),
- horizontal gene transfer ("transference"),
- erroneous homologization of structures which in reality are only superficially similar.

The first two sources of error can be identified with a detailed character analysis whenever the characters are complex enough. By way of contrast, transferred genes are true homologies, which however occur in non-homologous surroundings. Therefore, comparing different genes of the same organisms, gene trees will not be congruent. That superficial similarity is not recognized as such is

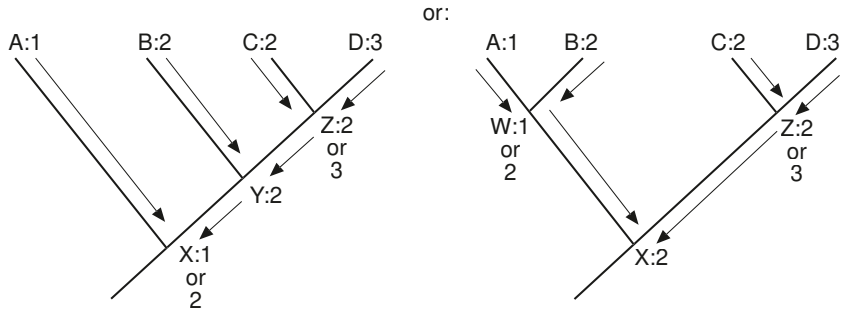


Fig. 131. Determination of "node characters" for the ground patterns of W, X, Y and Z. (1-3: transformation series of a character; arrows: direction of reconstruction).

generally a consequence of an inattentive or careless way of data acquisition.

Different cladistic indices are used to describe the numerical ratio of potential apomorphies to homoplasies (ch. 6.1.9.1).

6.1.6 Manipulation of the data matrix

As the structure of a dendrogram depends on the distribution of character states among taxa and also of the character weights in a data matrix, often the topology can easily be modified by small changes. "Disturbing" characters can be eliminated, "fitting" characters can get a higher weight, whereby frequently the unfounded "feeling" of an author is the basis of these decisions. Such manipulations can be unscientific or even fraudulent. Objective, scientifically justified manipulations can only be performed on the basis of a new evaluation of the quality of characters. It is discussed in ch. 5 how to estimate the quality of characters.

6.1.7 Cladistic reconstruction of ground patterns

The procedure to reconstruct ground patterns which satisfies the laws of phylogenetic systematics discovered by W. Hennig requires phenomenological character analyses (see ch. 5.3.2) preceding tree construction. With phenetic cladistics, ground patterns are not found during character search, but on the basis of character *distribution* in a dendrogram. This has the consequence that the arrangement of characters in ground patterns ("characters in inner nodes") becomes dependent of the topology of the complete dendrogram and is independent of the information content of the characters themselves.

Each cladistic determination of tree lengths (see MP-methods, ch. 6.1.2) makes assumptions on character states in inner nodes of the dendrogram, and thus on character states in ground patterns. Using Wagner parsimony with its assumption that characters are reversible, these states are determined with the following steps:

- construct a rooted or unrooted most parsimonious dendrogram.
- Select a pair of neighbouring terminal taxa. If these taxa are species groups they must be represented in the data matrix by ground pattern characters reconstructed in a previous analysis.
- Select a character.
- Select for this character the common state (majority rule) found in the ground pattern of the two terminal taxa and use it as the character state of the node joining the terminal taxa. If this cannot be determined because two states are equally common, choose the Boolean operator "or" for the two states of the terminal taxa.
- Determine the character state in the next "lower" node based on the states of the more distal nodes connected to it. The latter can be ground patterns or characters of a terminal species. Should information on one of the neighbouring distal nodes be lacking, first of all the character state in this neighbouring node has to be reconstructed with a separate analysis, starting with the terminal taxa which are connected to this node.

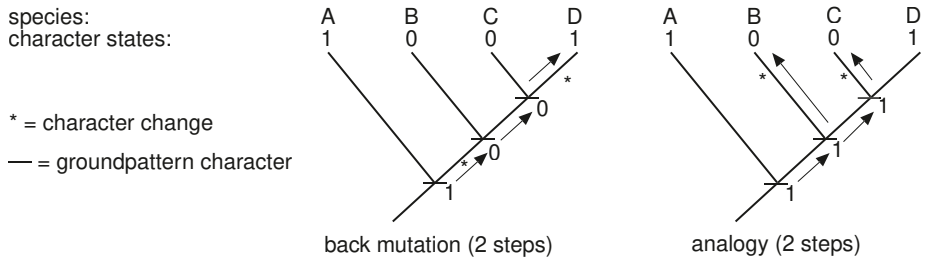


Fig. 132. Alternative assumptions for the reconstruction of ground patterns: favouring reversals (left, corresponds to the "Dollo parsimony") or analogies (right, corresponds to the "Camin-Sokal parsimony", compare Fig. 127).

The determination of node characters implies assumptions on character state changes. The uncertainty of states in some of the inner nodes, however, prevents an objective determination of the place (branch) where some of the characters change. For the calculation of character state changes one can therefore choose between alternative algorithms, which either favour reversals or analogies (parallelisms) for all characters (Fig. 132).

Depending on the algorithm one can obtain alternative ground pattern states. To chose between these alternatives, additional information is needed (e.g., on the probability that similarities may be analogies). In some computer programs these alternatives are implemented with the names DELTRAN ("delayed transformation": analogies are favoured) and ACCTRAN ("accelerated transformation": reversals are favoured).

The example (Fig. 132) shows that the number of character state changes detected and thus the "length" of a tree is independent of the chosen ground pattern-inferring algorithm, although the assumed course of evolution is different. Therefore cladistic computer programs do not require an unequivocal determination of character states in ground patterns (further statements on the reconstruction of ground patterns with exact Wagner algorithms in Swofford & Maddison 1987).

Ground patterns are inevitably reconstructed with cladistic computer programs when the character matrix contains polarized or irreversible characters or when characters of a stem-species ("inner node characters") are given. With some computer programs the user can get a list of character states or of potential apomorphies of stem species after the analysis. These ground patterns are calculated only from the distribution of character states in the given topology and are independent of the phenomenological estimation of the probability of homology or of the probability that a specific evolutionary process of character state change happened.

Note that in contrast to the method shown in Fig. 111, ground patterns estimated with popular parsimony methods will not show for which characters insufficient information is present. The reconstruction of ground patterns with the method of phenetic cladistics entails the **systematic error** that a deficiency in the available information is not reflected in the reconstructions. *The source of this uncertainty is the omission of character analyses*. For each character of the "inner nodes" of a dendrogram a state is given, no matter whether the available information is sufficient or not. When for example a sequence position is so variable that it does not unequivocally fit to a split, this position is nevertheless counted in parsimony methods in the same way as conserved positions. To prevent this, a weight has to be ascertained for each character in a separate analysis. For characters with ambiguous states in inner nodes a question mark can indicate the lack of information.

Many cladists do not understand how the quality of characters can be evaluated *a priori*. Therefore they call disdainfully the reconstruction of ground patterns using Hennig's method without computer programs "intuitive" (e.g., Yeates 1995). However, a combination of an evaluation of the probability of homology with a subsequent application of the maximum-parsimony method (ch. 6.1.2) agrees well with Hennig's method (phylogenetic cladistics) and is necessary to reduce the risk of using data of low value.

For quantitative characters which can change continuously, like the average body weight or the

immunological distance to an outgroup taxon, methods have been developed which allow the estimation of the state in the ground pattern. However, neither episodic changes of substitution rates nor the estimation of the probability of homology can be taken into account (see Maddison 1991).

When a species shows **polymorphic characters,** with some individuals bearing plesiomorphic and others apomorphic states, one must test whether

a) the polymorphism was already present in the last common ancestral population of this species and its sister species (polymorphism in the ground pattern);
b) if the apomorphic state is a novelty that evolved within the species and which does only occur in certain monophyletic populations;
c) or if the derived character states could have originated convergently several times.

Evidence for the presence of a polymorphism *in the ground pattern* (or in the last common ancestor population) is when the polymorphism occurs in the sister taxon or closely related taxa and at the same time analogy can be ruled out. Polymorphic characters in phylogeny reconstruction can only be evaluated when it is possible to homologize each morph and when the evolution of a morph can be reconstructed, as in the case of sexual dimorphisms.

The analysis of a polymorphic character yields a **gene tree** which often does not correspond to the species tree because the evolution of new alleles takes place prior to speciation. Therefore, several alleles can occur simultaneously in different species. The resulting problems correspond to those occurring during the analysis of paralogous genes (Fig. 6, 7).

Shared polymorphic characters in different species are to be expected especially in closely related species which have short divergence times. However, it can never be ruled out completely that several gene variants also coexist in populations over longer periods of time. The effect would be that gene trees do not correspond to the sequence of speciation events (Fig. 6). When analysing several genes which are inherited independently (this would not be true for mitochondrial genes, for example, which are duplicated as a unit), it is to be expected that trees for polymorphic genes differ: it is unlikely that the same gene tree evolved several times by chance. Congruence of the topology is with higher probability the result of the same historical processes. (This does not necessarily mean that the correct species tree has been found when congruence of several gene trees is obtained; see "plesiomorphy trap" in ch. 6.3.3).

6.1.8 Rooting of unpolarized dendrograms

Using MP-algorithms and taking a dataset with unpolarized characters, only unrooted trees can be constructed. Rooting is not necessary for the cladistic data analysis itself, but will for the interpretation of the course of evolution of characters and of populations. A determination of the polarity of the time axis in a dendrogram is possible with the following methods:

– using characters whose polarity has been determined in a previous character analysis and that are coded accordingly (ch. 5.3).
– Use of irreversible characters.
– Determination of at least one taxon as outgroup (before or after inference of the tree: see cladistic outgroup addition, ch. 5.3.3).
– Assignment of a character set to an inner node.

It should not be ignored that with the *a priori* determination of character state polarity additional valuable information flows into the analysis. Many alternative topologies can be dropped if they require a reversal of the polarity of important characters. The necessity for an *a priori* character analysis is especially advantageous because it forces the scientist to differentiate potential homologies and chance similarities.

Note that **midpoint rooting** is just the definition of a point in the middle of the longest path between terminal taxa found in a dendrogram. This would be the correct root only when the branch lengths are proportional to time or when they represent the correct number of substitutions of a sequence evolving like a perfect molecular clock. In practice, midpoint rooting will not be reliable in most cases.

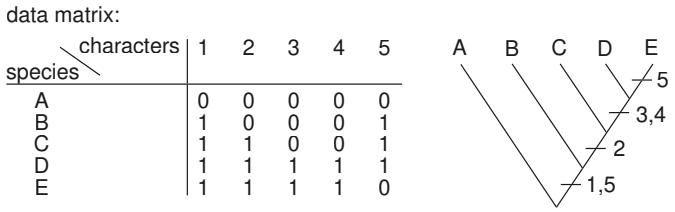


Fig. 133. Example for the calculation the consistency index. The total number of character state changes occurring in the characters is M = 5, the total number of steps in the dendrogram is S = 6. Therefore, CI = 5/6 = 0.833. (The value is often multiplied by 100: CI = 83.3). By addition of a trivial character which represents an autapomorphy of species E and hence has no influence on the topology, the CI-value improves: M = 6, S = 7, CI = 6/7 = 0.857.

6.1.9 Cladistic statistics and tests of reliability

One should never forget that the cladistic tests of congruence between the information in a dataset and the information in an inferred topology has purely methodological objectives, but they do not allow statements on the probability of *evolution* of characters or of groups of species. The claim that a dendrogram is trustworthy because cladistic tests yielded good values is unfounded. A good test value at best proves that the topology reflects well the information used by the method and contained in the dataset. A statement on the *quality of the dataset* cannot be obtained with these methods. Whoever is interested in the exploration of the quality of the data has to analyse the probability of homology (ch. 5.1) and the signal-noise-relation (ch. 6.5, 14.7).

6.1.9.1 Consistency index, retention-index, F-ratio

The **consistency index** evaluates the number of homoplasies as a portion of the total character state changes of a topology. It is calculated as follows:

The number of character states which are considered in a dataset for character **i** is $\mathbf{n}_i$. Then, the lowest number of character state changes $\boldsymbol{m}_i$ which are to be expected in a topology is $\boldsymbol{n_i - 1}$, implying a single occurrence of each apomorphic state.

When $\boldsymbol{s}_i$ is the number of character changes occurring in a topology, the consistency index for a character $\boldsymbol{i}$ is

$$c_i = m_i / s_i$$

If there are no homoplasies for a variable character in a given topology, we get $c_i = 1$, and for invariable characters $c_i = 0$. When homoplasies occur, we find $s_i > m_i$. The consistency index ***CI*** for the whole topology is calculated from the sum $\boldsymbol{M}$ of all $\boldsymbol{m}_i$ and the sum $\boldsymbol{S}$ of all character state changes $\boldsymbol{s}_i$ present in the topology :

$$CI = M / S$$

When homoplasies are present for a character in a topology, this character shows more state changes $\boldsymbol{s}_i$ than the minimum number of changes $\boldsymbol{m}_i$. The consequence is that the index decreases. If no homoplasies are present, we obtain CI = 1. With this test a comparison of datasets and of topologies is possible: the closer the CI-value is to 1, the better is the fit between topology and dataset.

However, with the same number of homoplasies the CI-value also depends of the number of taxa and characters, as well as on the presence of autapomorphies (Fig. 133). Therefore, from a purely methodological point of view it is not a good measure. For a bush topology which does not contain synapomorphies, the CI-value is greater than zero, a fact that does not comply with the original interpretation of the index. A character that shows two convergent trivial (autapomorphic) character states in a topology has the same CI-value as a character with convergent states of which one is a trivial autapomorphy and the other one a group-supporting synapomorphy. Only in the latter case the topology is at least partially supported. This weakness of the consistency index does not occur in the retention index.

The **homoplasy index *HI*** is complementary to the consistency index ($HI = 1 - CI$) and is taken to be a measure for the portion of character state

data matrix:

species \ characters	1	2	3	4	5
A	0	0	0	0	0
B	1	0	0	0	1
C	1	1	0	0	1
D	1	1	1	1	1
E	1	1	1	1	0
maximum number of steps	1	2	2	2	2

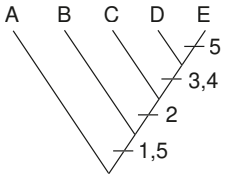


Fig. 134. Example for the calculation of the retention index. The maximum number of steps l_{max} is a sum of character states in a data matrix, whereby the rarer character state of each character is counted per column. For this example we get $l_{max} = 9$, and as in Fig. 133 M = 5 and S = 6. Hence follows: RI = (9-6) / (9-5) = 0.75.

changes which are caused by homoplasies. In principle, it could be a measure for the noise present in data. However, the weakness is the same as in the consistency index.

The **retention index** (***RI,*** see Farris 1989a,b) was designed to be a measure for the amount of putative synapomorphies (in relation to a given dataset) which are retained in a topology. The more putative analogies occur, the lower is the RI-value. To achieve this effect it is tested how many homoplasies occur in the topology, and the number is put into relation to the maximum number of possible homoplasies. The number of symplesiomorphies retained (conserved) in a topology is counted as complement of the number of homoplasies. Since symplesiomorphies and autapomorphies do not occur in form of homoplasies, they do not alter the index value, an important difference to the consistency index. The index is calculated as follows:

The length of the given dendrogram (number of observed character state changes in a topology, including autapomorphies and analogies) is ***S***, l_{max} is the maximal possible length of a dendrogram for a given dataset, ***M*** is the sum of the character state changes m_i of each character**,** added for all characters. The **retention index** *RI* is calculated as follows:

$$RI = \frac{l_{\max} - S}{l_{\max} - M}$$

The value l_{max} is obtained from the sum of the values l_i for all characters ***i*** of a dataset. l_i corresponds to the number of character state changes of character ***i*** in a "bush diagram" (see Fig. 53), with the most common character state in the center (this is equivalent to the number of terminal taxa which show the less common character states). Therefore, the numerator is highest when the characters do not show homoplasies and are distributed unambiguously, like synapomorphies. The value is lowest when the character states of different taxa are analogies, and not synapomorphies. The denominator is constant for a given dataset. (A simultaneously developed homoplasy-index (**homoplasy excess ratio**: Archie 1989, Farris 1991) is equivalent to the retention index).

The **F-ratio** (F for "fidelity") also serves the quantitative description of the homoplasy content in a topology for a given dataset. It is obtained by converting the data matrix into a matrix of phenetic (uncorrected) distances, which shows the number of character state differences between pairs of species. This matrix is compared to a patristic distance matrix consisting of the number of steps on the path between species in the selected dendrogram. When no homoplasies are present, both matrices are identical (more details in appendix 14.10). This index also varies with the number of autapomorphies of terminal taxa; therefore it cannot be recommended.

A further indication for the number of homoplasies causing "noise" in a dataset is the skewness of the distribution of tree lengths (see 6.1.9.3, 14.9).

In practice, these indices prove to be of minor importance because they do not allow statements on the quality of individual characters or alignments and they do not estimate the plausibility of hypotheses of monophyly.

6.1.9.2 Resampling tests

Bootstrap-Test

Often "bootstrap values" are used as an indication for the trustworthiness of hypotheses of monophyly (Felsenstein 1985). These numbers show the percentage of trees that recovered a putative monophylum (an inner branch) when trees are calculated from resampling of the original dataset (probability of recovery of a branch). Tests of this sort are used because they visualize the effect of homoplasies that support alternative topologies without the necessity to illustrate graphically all alternatives.

The following method is also called "non-parametric bootstrapping", because we want to test the reliability of the dataset in relation to an optimized topology without questioning the model parameters. The resampling consists of a random selection of characters (columns) from a set of data to assemble a new dataset of the same size (with the same total number of characters). In this new dataset some characters are naturally missing, whereas others occur twice or more. For each dataset the optimal topology (in MP-methods the "shortest tree") is calculated. When these steps are repeated for example 100-, 500-, or 1000 times one can state how often a putative monophylum occurs in these iterations. A given branch will only appear frequently in reconstructed trees when it is supported by several characters and when there are few characters for incompatible clades. The portion of cases with a given branch is usually expressed in percent. This percentage is called "bootstrap value". With a specific composition of a dataset it may happen that only those monophyla get values over 95 % which are, for example, supported by more than three potential apomorphies. Of course, if the characters are not weighted the result is independent of the quality of putative apomorphies. Many cladists think a value over 75-80 % corresponds to a high probability of monophyly. This, however, is a fallacy due to the following reasons:

a) with the bootstrap value neither the quality of the characters used nor

b) the quality of the taxon sample is estimated.

c) The bootstrap value cannot help to recognize in a reliable way the lack of information or the accumulation of random similarities in polyphyletic assemblages of species (Fig. 135).

The bootstrap value rather depends on the method of tree inference and on the number and distribution of characters in the data matrix and thus mainly allows a statement on the congruence between topologies and the structure of the data. Bootstrapping is, however, the most frequently used and one of the most intuitive methods of data quality evaluation.

High bootstrap values are obtained for a group of species in the following cases:

- A group is represented in a dataset with **many putative apomorphies** and has more supporting characters than alternative groupings and therefore appears repeatedly despite of the loss of several characters in different resampled datasets.
- Support of a group is based on **few apomorphies**, there are however no characters favouring alternative incompatible groupings, with the effect that only a single supporting character is sufficient to recover the group.
- A false grouping is supported as putative monophylum when there are no or only few apomorphies for the real monophylum (or for several real monophyla), but the dataset contains instead **several analogies** which separate a group that in reality is not monophyletic.
- High bootstrap values will also be obtained when **plesiomorphies** support a paraphyletic group while the real and incompatible monophylum is weakly supported, or when plesiomorphies do not occur in the selected outgroups. This can be a form of long-branch effect (ch. 6.3.3) or a problem caused by insufficient taxon sampling.

Due to these phenomena the resampling test indeed serves an important purpose whenever informative data are used, however, it is **not absolutely reliable in practice**, especially when working with DNA-sequences, which may show many analogies and unnoticed symplesiomorphies.

Consider a dataset with the four taxa A-D and with characters, half of which support the group (A, B), whereas the other half fits to (A,C). For each group a bootstrap value of 50 % is obtained

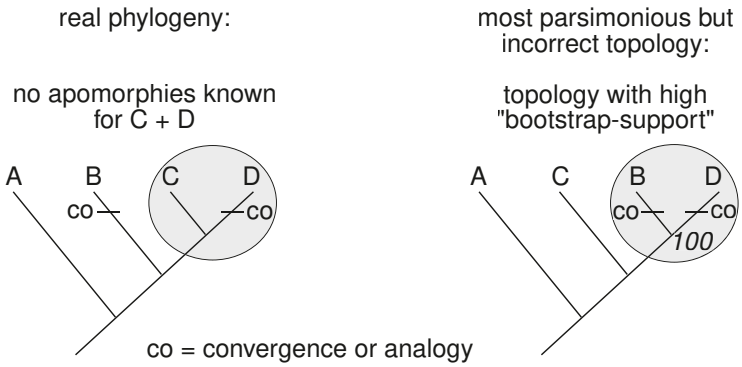


Fig. 135. Situation in which high bootstrap values support the wrong monophylum (B+D), when for the correct one (C+D) no characters are present in the data matrix.

and the consensus topology will show no resolution. To support unequivocally the monophyly of a group, the value should be greater than 50 %. Empirical observations, however, prove that even higher bootstrap values should not inspire confidence in the quality of the data and the monophyly of a group, which is not difficult to understand in view of the above-mentioned sources of errors.

Jackknifing (elimination test)

Bootstrapping in combination with maximum parsimony algorithms and re-grouping of branches ("branch-swapping", see ch. 14.2.1) are very time-consuming procedures, especially when large datasets are analysed. An effective alternative are elimination tests ("jackknifing"), which with adequate programming allow short computation times (Farris et al. 1996). The differences between the recovery values for clades obtained with both methods are small. For the jackknifing-test, randomly selected characters are eliminated from the dataset to calculate the most parsimonious topology for the remaining characters. These steps are repeated, for example, 1000 times. The frequency ***G*** for the recovery of a group of species is counted for all groups that appear in trees. Each character is eliminated with the same probability ***p***. In theory, if there are no question marks (unknown character states) in the dataset and if there is no conflicting evidence, the expected frequency ***G*** of a group depends on the number ***r*** of unambiguously supporting characters for a group (= potential apomorphies which do not show homoplasies): $G = 1 - p^r$ (Farris et al. 1996). Therefore the frequency is independent of the absolute number of characters and taxa. This formula, however, does not really explain the occurrence of a clade in a topology, because the most parsimonious topology only shows mutually compatible clades and the appearance of a clade depends on the number of characters supporting incompatible groups.

In the program JAC of Farris et al. 1996, the probability p for the elimination of characters is standardized to the value e^{-1} (= 0.3679). Higher values do not give a convenient relation between the number of supporting characters and the frequency G.

Parametric bootstrapping

"**Parametric bootstrapping**" is based on a Monte-Carlo simulation (ch. 8.4) and used for molecular data. A topology, which is assumed to be the optimal one, is taken in order to produce an artificial set of characters with the same length L as the original alignment. The evolution of an artificial sequence of the length L with randomly selected characters is simulated along the given topology with the help of a selected model of sequence evolution (software: see for example Rambaut & Grassly 1997). The result should be a set of artificial sequences with the same number of sequences and of sequence position as the original alignment, i.e. the model does not include insertions and deletions. The procedure can be repeated with other model parameters. The following steps have to be performed:

- Select a model of sequence evolution. Reconstruct a topology or choose a topology which shall be tested.
- Estimate model parameters using the maximum likelihood method (branch length, topology, ratio of transversions to transitions etc.) on the basis of the real data (compare ch. 8.3, 8.4, 14.6, selection of models with likelihood ratio test in ch. 8.1).
- Produce a new artificial data matrix (alignment) of the same length as in the original alignment given these model parameters and the topology to be tested.
- Calculate a new topology on the basis of the simulated data matrix.

With multiple repetitions of these steps it can be tested whether using the model assumptions again and again the same topology is found with simulated data. Thus it can be tested whether the evolution of a sequence is adequately described with the model (see Adell & Dopazo 1994), or if, for example, the branch lengths obtained in the simulation differ much from the given topology applying a molecular clock model. The method has been used to detect paraphyla which are produced by chance similarities in long branches, to analyse the influence of sequence length, of individual model parameters and of methods of reconstruction on the recovery of a topology. Further details and literature in Huelsenbeck et al. (1996b).

Bremer support

As an alternative to the bootstrap method, Bremer (1988) tested indirectly how many putative apomorphies support a monophylum in maximum parsimony analyses. Comparing the most parsimonious consensus tree from an MP-analysis with longer topologies, with increasing tree length more and more equally long topologies are found and thus one gets for longer trees consensus topologies in which only few or in the worst case no monophyla are distinguishable. The "Bremer support" or "**decay index**" indicates how many extra steps a consensus topology has to be longer than the most parsimonious one to collapse a branch (stem lineage) that was present in the shortest tree. This additional number of steps depends on the number of homoplasies. With the same quantity of supporting characters (= "branch length of the stem lineage"), a monophylum which is distinguished by unique apomorphies has a higher "Bremer support" than a monophylum whose supporting characters also occur in other branches. The analysis only makes sense when characters are weighted according to their probability of homology.

For example, the smallest number of expected character state changes can be counted for a given dataset (sum of all steps between character states considered in a data matrix, corresponds to *M* in the consensus index, see ch. 6.1.9.1). Assuming that the shortest consensus topology reconstructed for this dataset shows 151 steps and the smallest expected number of steps is 88, then 151-88 = 63 steps have to be attributed to homoplasies. For a higher tree length there are more equally long topologies with more homoplasies and fewer putative monophyletic clades. If, for example, a clade is present in the shortest tree (151 steps), but not in the consensus for the second shortest tree (152 steps), the value "1" is declared as *decay index* or *Bremer support* for this clade. This means that the omission of a single (specific) character in the dataset can already have the effect that the monophylum is not recovered.

To calculate the Bremer support is time-consuming, because all trees have to be found that are 1 step away from the strict consensus tree, and then those that are two steps away, etc., to get the consensus of the near minimal length trees. Furthermore, it is not clear when a support value really is reliable. For example, closely related species might share only one or a few apomorphic character states and get a low Bremer support, even if the support is significant in cases when there is no contradiction in the dataset (for software see Sorenson 1999).

6.1.9.3 Distribution of tree lengths, randomization tests

The fundamental idea of these tests is the comparison of an optimal result with a random distribution of results obtained on the basis of the same data or from randomized data. The optimal result can be a shortest tree topology, for example. A clear deviation from the random distribution of tree lengths indicates the presence of "real" (non-random) signal, the optimal tree would then

reconstruct at least partially the historical processes and contain at least some of the true phylogenetic relationships of taxa.

A permutation test can consist of the repeated assemblage of an artificial set of characters which contains the same number of taxa and of character states as the original dataset. The characters are however distributed randomly among the taxa. Dendrograms are constructed with the same method that has been used for the original data. In the **PTP-test** (permutation-tail probability test) the portion of taxa which have a specific character state for a character remains constant, but the taxa are selected by chance. Such random datasets can be produced one hundred times, for example, so that after calculation of the shortest topologies with the MP-method a length distribution for all shortest dendrograms is obtained. If the optimal topology estimated on the basis of the original data has the same length as say the 5 shortest of 100 random topologies, the conclusion is that with a probability of 95 % the estimated topology contains more elements of the real order of taxa than can be obtained by chance alone.

One should not forget, however, that no statement is possible on

- the support of individual nodes,
- the adequacy of the tree construction method, because the same method has been used for the production of all test results (topologies for random data),
- the quality of the real dataset, because the accumulation of chance similarities, the presence of symplesiomorphies (due to insufficient taxon sampling), or of convergences in real data can produce well supported groups, which, however, are not monophyla,
- the quality of the selected sample of species (see ch. 6.3.3).

Therefore the utility of these randomization tests is restricted (further details in ch. 14.9).

6.1.10 Can homologies be identified with the MP-method?

It has already been discussed what homologies are and how they can be recognized with a phenomenological approach (ch. 4.2 and ch. 5). Pattern cladists have claimed that homologies can be identified using an MP-algorithm and excluding any influence of the subjectively judging scientist. Knowing that a dendrogram implies statements on character states in ground patterns of monophyla, it is erroneously deduced by many cladists that homologies are found in topologies and that the evaluation of hypotheses of homology *prior* to the MP-analysis can be completely abandoned.

Relying on this cladistic reasoning it is overlooked that one runs into a **circular argument** when homologies are identified *a posteriori* (see ch. 6.3.1 and Fig. 139). The MP-method needs the axiomatic assumption that characters of the same weight are homologies with a similar probability.

Already when assembling a data matrix, hypotheses of homology *for the frame homologies* are implied for each column (with species in rows), which are not tested in the further course of the analysis. The cladistic congruence test (see ch. 5.1.1) can only serve to test the compatibility of *detail homologies* (character states) in the optimal tree. If, for example, you code different eye colours, the homology of the eye itself is not in question. The only characters that are tested for correspondence are the colours, the event that is tested for uniqueness is the change of eye colour.

Example: after a cladistic analysis of characters of the Metazoa, Schram (1991) obtained a phylogenetic tree in which Tracheata (insects and myriapods) and Onychophora (velvet worms) are sister groups. The most parsimonious topology contains a single synapomorphy supporting this group, the "whole-limb mandible" with the states "present" and "absent". Taking the cladistic analysis as a test for congruence, it has to be concluded that the "whole-limb mandible" is a homology in Tracheata and Onychophora. But this conclusion is a **circular argument**, because the assumption of homology has already been coded in the data matrix. Without this coding this sistergroup relationship would not exist. The same holds for the character "tendency to develop tracheae", which has been coded as apomorphic character of Tracheata, Onychophora, and Chelicerata. This "tendency" is not a material character, because there exists no common genetic coding for it (is a tendency a gene?). The criterion of congruence is not suited to identify characters which are not homologies.

The same problem is known from sequence analysis: the quality of an alignment in the sense of the probability that a position has been homolo-

characters / organisms:	1 (forest spider)	2 (cave spider)	3 (cave cricket)	4 (crayfish)
body white, not pigmented	—	+	+	—
eyes large, functional	+	—		+
legs elongated, slender	—	+	+	—
spinnerets on caudal body	+	+	—	—
hair sensilla elongated	—	+	+	—

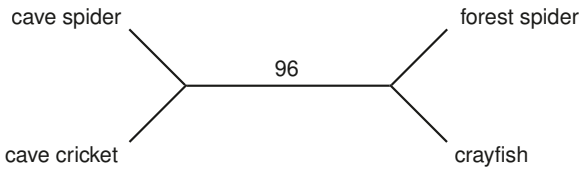


Fig. 136. Fictitious character table for the comparison of cave arthropods and epigean species (+: character present; –: character absent). Below: result of an analysis of the matrix with the program PAUP (MP-methods with the following settings: 500 bootstrap runs, TBR-branch swapping). In the unpolarized diagram (no outgroup has been determined) in each case the two spiders are not closely related to each other. Either the cave animals belong to a monophylum or the forest spider and the crayfish. This split is found in 96 % of the 500 bootstrap-runs (for the method see ch. 6.1.9.2).

gized correctly is not tested with the reconstruction of the phylogenetic tree. The alignment position is a *frame character*. The alignment process is therefore an independent step prior to the phylogenetic analysis. It has to be remembered that cluster analyses or parsimony analyses can be performed with any data, even with non biological patterns such as components of minerals. Also random data may produce tree topologies. The fact alone that a dataset can be projected on a tree topology is not a proof for the existence of biological homologies.

At this point a fictitious, transparent example shall demonstrate how a cladistic identification of homologies can produce errors. Methodically, this example corresponds to an exact, "automated" cladistic analysis of the phenetic variant:

Let us assume that an untrained zoologist detects a number of arthropods in a cave. He compares these animals with organisms living outside the cave and takes notes on those characters which attract his attention. In Fig. 136 the name of the organisms not known to our cave researcher is shown in brackets.

The result of this analysis (Fig. 136) is methodically incontestable from the point of view of a pattern cladist, even though one would desire that more characters are coded. We get a group of cave species separated from those species living overground. The characters of the cave species are derived, and therefore it can be postulated that this group is monophyletic. This implies that probably the adaptations to living in a cave (depigmentation, reduction of eyes, increase of appendage length and elongation of hair sensilla) are homologies. The spinnerets are not homologous. The trained systematist, however, recognizes that the mistake results from the use of many similarities which have a small probability of being homologies and there is little complexity in these modifications. Only one highly probable and complex homology is contained in the data matrix (the spinnerets with their silk glands and specialized muscles). The effect of this character is suppressed in the MP-analysis by the numerical predominance of characters of low value: the group {cave spider + cave cricket} is united by four characters, whereas only one character supports the monophyly of spiders. **With the MP-analysis the real homology is not detected!** The

erroneous topology could only be avoided by evaluating characters and their probability of homology *prior* to the analysis.

One could object to this fictional example (Fig. 136) that it is not realistic because in practice the systematist would consider more characters. However, one cannot count on the diligence of scientists as is visible in the above-mentioned example (analysis by Schram 1991) and in many similar cases. With the criterion of congruence it cannot be tested whether a scientist has worked hard (see "quality of the receiver" in ch. 1.4.5). Furthermore, quantity of characters is not the same as quality. And, more seriously, usually only a few potential apomorphies are found when comparing closely related species. Their value has to be tested independently of tree construction (compare ch. 5.1). Numerous published studies containing the same mistake, unnoticed by the authors, are equivalent to this fictional example. The following consideration can clarify which consequences the cladistic homologization has: should a dataset be composed so unfavourably that the topology derived from it shows polyphyletic Aves, it must be concluded that the character "feather" is not homologous! To accept this conclusion it is really necessary to find convincing characters of great weight that support a tree with polyphyletic birds.

The ***a posteriori* determination of homologies**, i.e. the formulation of a hypothesis of homology on the basis of a phylogenetic tree, is only convenient when the tree has been built with other characters of high weight (with "valuable homologies", see ch. 5.1). Only in this case the procedure is not circular. In practice one will only resort to this form of homologization when a character is not complex enough to infer for it a high probability of homology.

Examples: the eyes of many hypogean or deep-sea crustaceans are reduced (Fig. 94). It appears that the reduction does not require a specific and complex series of mutations, and the reduction obviously offers a selective advantage (saving of energy and material). The character " eyes reduced " has therefore little weight and as a single character generally is not suited to substantiate the monophyly of a group. – The blind crustaceans of the taxon Microcerberidae (Isopoda) live subterraneously in ground water. They are identified as members of a monophylum because they show several evolutionary novelties such as special mouthparts and simplified but specifically shaped pleopods. As monophyly of this group is well founded, the lack of eyes can also be considered as a character of the ground pattern of this taxon: the homology "eyes reduced" can only be determined *a posteriori*. This, however, is the application of the *criterion of compatibility*, not the *criterion of congruence* (ch. 5.1).

The criterion of congruence actually is significant for the test of hypotheses on the homology of apomorphies that were postulated and substantiated *a priori* and imply hypotheses of monophyly (compare Fig. 10). However, substantiation and test are two steps that are logically independent of each other.

6.1.11 Sources of errors of phenetic cladistics

Since the result of a maximum parsimony analysis is already determined with the composition of a data matrix and the weights applied to characters, the scientific substantiation of hypotheses, i.e. the steps that are decisive for the result, takes place at the level of the selection of taxa and characters and with the coding and weighting of characters (see Figs. 137, 138). Analyses of published phenetic cladistic studies prove that the mistakes listed below really occur in practice (compare Wägele 1994a):

- Characters coded for species are based on incorrect or superficial observation of organisms, they do not exist in nature.
- Characters selected for supraspecific taxa are based on erroneous reconstruction of ground patterns of these taxa or they represent characters of derived species.
- Important homologies that are already known are ignored.
- Apomorphies are defined for the ingroup which in reality are most likely plesiomorphies.
- The selected characters have only a small probability of homology, the patterns recognizable in the data matrix are based on "noise" (homoplasies). The selected characters are most probably analogies or convergences.

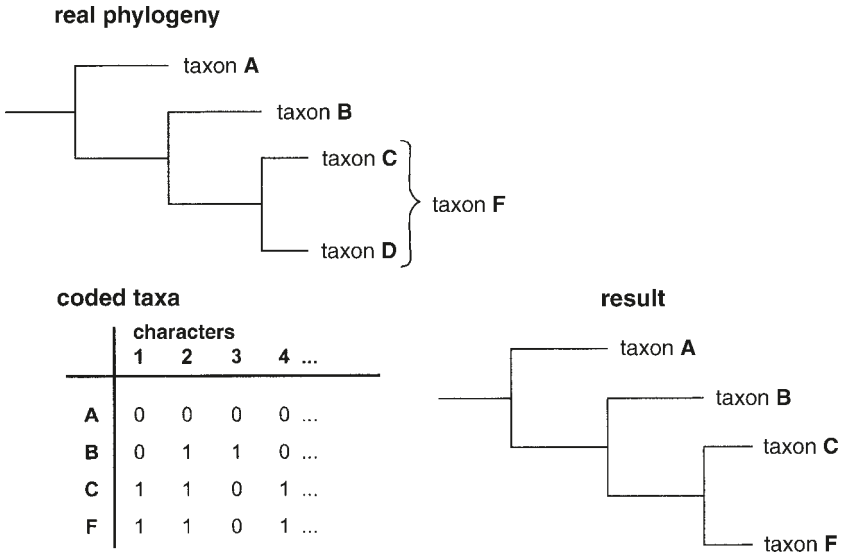


Fig. 137. In the true phylogenetic tree, the taxa C and D together constitute the taxon F. When C and F are coded separately in a data matrix, a wrong sistergroup relationship of C and F is obtained.

- Weights are not proportional to the estimated rank of probabilities of homology.
- Wagner-parsimony allows reversion (repeated independent origin) of complex characters.
- The selected outgroup in reality is part of the ingroup.
- The selected outgroup shows only few plesiomorphic character states, wherefore polarity of most characters determined with the cladistic outgroup addition is not correct.
- Characters were only studied in few outgroup taxa, wherefore it is overlooked that apparent apomorphies of the ingroup also occur in other taxa.
- Terminal taxa are not monophyletic.
- Encaptic terminal taxa are coded separately (e.g., the taxa Marsupialia and Mammalia appear in the data matrix next to each other). In this case it will not be detected that one taxon is part of the other one (Fig. 137).

Besides these mistakes, which are made during the assemblage and analysis of the data matrix, another deficiency of many published cladistic analyses is the omission of a test for the plausibility of the obtained dendrograms. It always has to be asked which consequences a hypothesis of phylogeny has on the presumed evolution of ways of living, of physiological abilities and on the implied evolution of the anatomy of organisms (see ch. 10).

Some authors state that maximum parsimony requires an improbable evolutionary scenario that involves the fewest number of changes. This is a misunderstanding: constructing a tree the MP method does not search for short cuts of evolution (e.g., directly from fish to whale instead of taking the detour via terrestrial tetrapods), but it minimizes the number of convergences, which is a sound probabilistic approach (see ch. 5.1).

6.2 Hennig's method (phylogenetic cladistics)

W. Hennig (1913–1976) has the merit to have developed a method of systematics which is strictly founded on the logic of scientific argumentation in the sense of Karl Popper. Hypotheses on homologies and on monophyla can be substantiated and falsified with intersubjectively verifiable criteria.

It has become a habit in English speaking countries to apply the term "cladistics" to variations of the MP-method (ch. 6.1), which originally did not root in the thoughts of Hennig, but also to Hennig's own method. This is understandable, because the deductive step used for tree construction, i.e. the application of the principle of parsimony to select shortest topologies, is the same in both approaches. However, the methodologies of the analyses are very different. Therefore in German the term "Kladistik" specifically refers to *phenetic cladistics*. The method of "phylogenetic cladistics" became popular with the methodically precise studies of W. Hennig. The major principles, however, have been recognized and used previously by other authors (Craw 1992). Hennig's method has been developed for morphological characters, but it can also be applied to other discrete characters including DNA sequences.

The graphical illustration of the diversification of life with the help of tree graphs and also the metaphor "tree" became popular with the theory of evolution. Darwin (1859) illustrated the extinction of species and successive speciations with a tree graph, but he did not develop phylogenetic trees. Already in 1864 Fritz Müller applied outgroup comparison and ontogenetic criterion and he pointed out evolutionary novelties. The Russian paleontologist Woldemar Kowalewski also used evolutionary novelties to differentiate groups of species and he reconstructed ground patterns for groups such as the ancestors of ungulates. These approaches, however, lacked a clearly formulated methodology and the application of methodical principles was inconsequent. T. J. Parker (from New Zealand) had already mapped in 1883 characters on a phylogenetic tree of rock lobsters. It also had been clearly recognized by A. Neef (1919) that a classification system for organisms should correspond to the phylogenetic tree and that this is an illustration of a series of "splittings of species". In Italy, Daniele Rosa, a specialist for annelids, discussed in 1918 essential principles of cladistics (including amongst others the monophyly of taxa, avoidance of paraphyletic groups, the end of a species after a speciation). However, it is not known whether W. Hennig knew of Rosa's work. Other authors, (e.g., E. Meyrick, A. Dendy and J. W. Tutt) used some aspects of phylogenetic systematics in their work were. Konrad Lorenz (1941) introduced an argumentation scheme prior to Hennig. Many scientists contributed to establishing the relationship between phylogeny and systematics and to develop the necessary methods. However, Hennig in his analyses of insect phylogeny worked out very clearly and convincingly the significance and advantages of his phylogenetic argumentation. It is due to the influence of these studies that the method and the terminology were accepted and propagated by an important part of the scientific community. The application of the principle of parsimony to the reconstruction of phylogenetic trees, which in the eyes of many contemporaneous systematists is the heart of cladistics, was not mentioned by Hennig in his influential book (1966) in form of an instruction for data analyses. This principle is rather implied between the lines and was first specified explicitly by cladists (e.g., Kluge & Farris 1969, Farris 1970) and later by authors describing Hennig's approach in a more concise way (e.g., Ax 1984, Ax 1988).

In his first important book on the theory of phylogenetic systematics, Hennig (1950) explains among others how the boundaries of species have to be defined along the time axis, that supraspecific taxa only have a relation to reality when they are monophyletic, and that monophyla comprise the last common stem-species as well as all the descendants of the last stem-species. Hennig published the term "paraphyletic" only in 1966. In 1950, Hennig had not yet introduced the tools necessary for the identification of monophyla we know now, but he had chosen the right way which led to later improvements. He perceived the significance of Haeckel's biogenetic law for the determination of character polarity ("criterion of the ontogenetic precedence of characters") and the criterion of character complexity for the evaluation of the probability of homology. Hennig at first introduced the terms "apomorphic" and "plesiomorphic" for taxa and not for characters (Hennig 1949), a usage that is methodologically not convenient. The present significance of these terms was elucidated by Hennig (1953) together with the introduction of the prefixes aut-, syn- and sym-. The much read English version of his 1950 book, published in 1966, contains several improved definitions and explanations, for example on ways to determine character polarity (see Richter & Meier 1994).

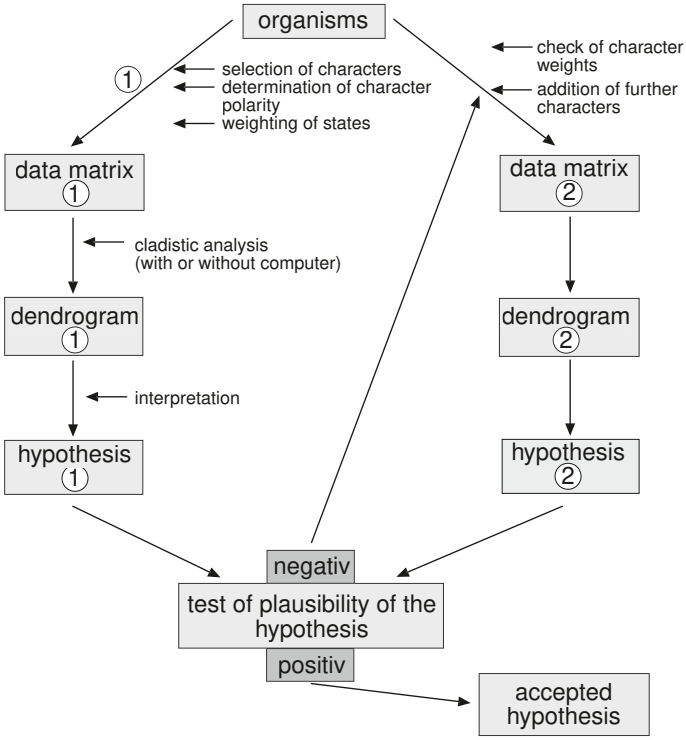


Fig. 138. Flowchart for a phylogenetic analysis with Hennig's method. By contrast, pattern cladists often concentrate only on the step "cladistic analysis".

A modern phylogenetic analysis in the sense of Hennig requires the following steps (Fig. 138):

- Search for similarities which could be used as characters within the analysed species groups.
- Perform a phenomenological character analysis (see ch. 5) to differentiate between apomorphies, plesiomorphies and analogies. To do so, characters have to be homologized in a first step (ch. 5.2) and characters of low probability of homology will be rejected. Character transformations should be weighted according to their estimated probability of homology. Character polarity has to be determined *a priori* for the selected homologies by outgroup comparison (ch. 5.3). A cladistic determination of polarity by outgroup addition is avoided (ch. 5.3.3).
- Apomorphies (or, more accurate, hypotheses on apomorphies) substantiate hypotheses of monophyly.
- Synapomorphies are evidence for sistergroup relationships (Fig. 73, 78, 124).
- Monophyla which serve as terminal taxa are only represented by reconstructed ground patterns. The reconstruction of ground patterns is carried out with a character analysis (ch. 5.3.2).
- When incompatible monophyla are found, those groups are retained which are supported by the larger number of weighted homologies and which are only compatible with the whole topology. This total topology is at the same time the "most parsimonious" one. The search for such topologies can be done with the popular MP-algorithms or 'by hand'.
- The most parsimonious topology (or the topologies) is (are) the basis for the reconstruction of evolution, for hypotheses on phylogeography and the historical dispersal of a group and for the description of evolutionary scenarios.

In this way the dendrogram is obtained successively by the identification of single monophyla. The result must not contain incompatible groups (the corresponding Venn diagram has no intersections), all monophyla should fit to an encaptic

order. Incompatible groupings (see ch. 3.2.2) indicate errors in character analysis such as inclusion of uninformative characters or of unrecognized convergences (sources of errors: ch. 6.1.11).

Hennig did not give special emphasis to the estimation of the probability of character homology (weighting) and did not explain its importance, wherefore the significance of the evaluation of hypotheses of homology has been overlooked in the cladistic literature. But already in 1950, for example, Hennig discussed the phylogenetic interpretation of intraspecific polymorphisms, rejected the equation of "group of common descent" ("Abstammungsgemeinschaft") with "community sharing similarities" ("Ähnlichkeitsgemeinschaft"), he insisted on the "evaluation and differentiation of individual characters" in connection with allometries noted in interspecific comparisons. In the chapter "the rules for the evaluation of individual morphological characters etc." he describes criteria of homology, for example the criterion of character complexity. He also uses the concept of weighting of homologies, even though he never weighted numerically.

When computer programs are used for an application of the MP-method of tree reconstruction (ch. 6.1), the following conditions and steps have to be taken into account:

- the terminal taxa of the data matrix have to be monophyletic and must be represented by ground pattern characters of high probability of homology within the terminal monophylum.
- Moreover character polarity has to be determined whenever possible by phylogenetic outgroup comparison before the cladistic step of the analysis (ch. 5.3). Polarized characters should not to be coded as reversible (see handbooks of cladistic computer programs). The extent of the support for single monophyla can be estimated with cladistic tests (bootstrapping, Bremer support: ch. 6.1.9.2)

Dendrogram and data matrix form together an argumentation scheme from which it can be derived which apomorphies are implicitly used as evidence for the monophyly of individual groups. Publications of results should not only contain plates with trees and tree statistics but also a discussion of the arguments in favour of the homology and polarity of the apomorphies contained in the data matrix.

In any case the analysis is not concluded with the reconstruction of a dendrogram:

- with all available additional data it has to be tested whether the result is plausible (ch. 10).

In contrast to phenetic cladistics, Hennig's method (phylogenetic cladistics) allows an epistemologically well founded hypothetico-deductive analysis (Bryant 1989; see also Fig. 10).

6.2.1 Comparison of phenetic and phylogenetic cladistics

Although in phylogenetic systematics the cladistic method of tree construction (ch. 6.1) can be used as one of the available tools, it has a different function than in phenetic cladistics.

The aim of each **phylogenetic analysis** is the reconstruction of a dendrogram which can be accepted as a well corroborated hypothesis for the "phylogenetic tree". The information required for this are apomorphic homologies. To achieve this

a) individual organisms as representatives for species or for higher (supraspecific) taxa and
b) properties of individual organisms have to serve as samples representing characters of species.

The quality of the reconstruction naturally depends on the quality of the samples, wherefore this quality is checked prior to the reconstruction of a phylogenetic tree. The principles of phenomenological character analysis, of polarity determination, and of the delimitation of taxa described in chapters 4 and 5 serve this purpose.

The aim of an analysis performed with the methods of phenetic cladistics is also the construction of a dendrogram. However, the information used are the similarities compiled in character tables. Typically, the quality of characters is determined *a posteriori*, i.e. after the selection of a dendrogram. According to the view of pattern cladists, the maximum parsimony method serves the identification of homologies (for a critique of this attitude see ch. 6.1.10). The difference between the two approaches is illustrated in Fig. 139.

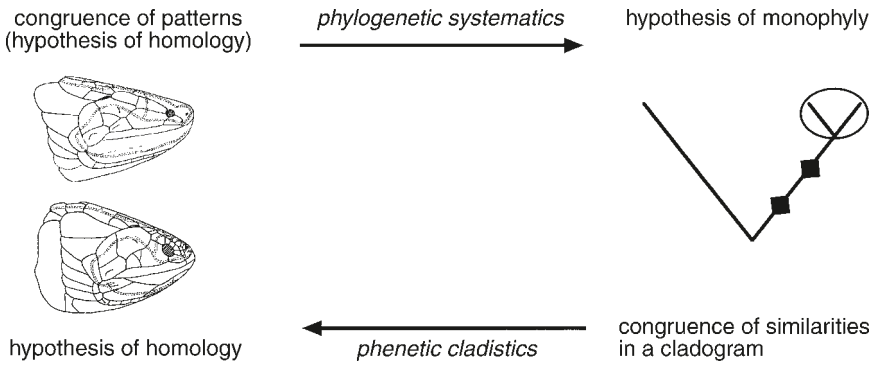


Fig. 139. Illustration of the methodological difference between phylogenetic systematics (phylogenetic cladistics) and phenetic cladistics.

In phylogenetic systematics the "**principle of reciprocal illumination**" is often used for the same argument which imprints phenetic cladistics: when two functionally independent characters support the same topology, they reinforce each other. As already explained, this argumentation is only convenient when (1) the probabilities of homology of all characters are comparable or when the characters were weighted accordingly and (2) when the topology is the most parsimonious one.

6.3 Cladistic analysis of DNA-sequences

For the cladistic analysis of DNA-sequences, the positional homology first of all has to be determined with an alignment method (ch. 5.2.2.1). The positions are the frame homologies while the patterns of specific nucleotides of single sequences occurring in an alignment are the character states evaluated during tree inference. As character state polarity is usually not determined, it is necessary to use algorithms for unordered characters.

An MP-analysis for unordered and equally weighted characters can be performed in the same way as for morphological characters. Since, however, often an unequal distribution of bases indicates that some substitutions occur more frequently than others, character transformations can be weighted differentially. *A posteriori* weighting should be avoided, because it leads to a circular argument (Ch. 6.1.10). There are two alternatives for the **a priori weighting** of character transformations (s. ch. 5.1):

- phenomenological weighting according to the contribution of single alignment nucleotides to the signal/noise ratio in the alignment (evaluation of the probability of cognition for homologies).
- model-dependent weighting (evaluation of the probability of events).

Methods for the latter approach dominated in cladistics for a long time, because there existed no concepts for the consideration of the complexity of nucleotide patterns. The phenomenological evaluation of the signal to noise ratio can now be achieved with the analysis of spectra (ch. 6.5). Further methods for the weighting of nucleotides according to their contribution to signal-like patterns are currently being developed. They rely for example on low weighting of particularly variable positions (Lopez et al. 1999).

6.3.1 Model-dependent weighting

With a model-dependent weighting of characters one leaves the purely phenomenological method. Nevertheless it shall be discussed here, because it is sometimes used in combination with cladistic analyses. Substitution models were developed for distance and maximum likelihood

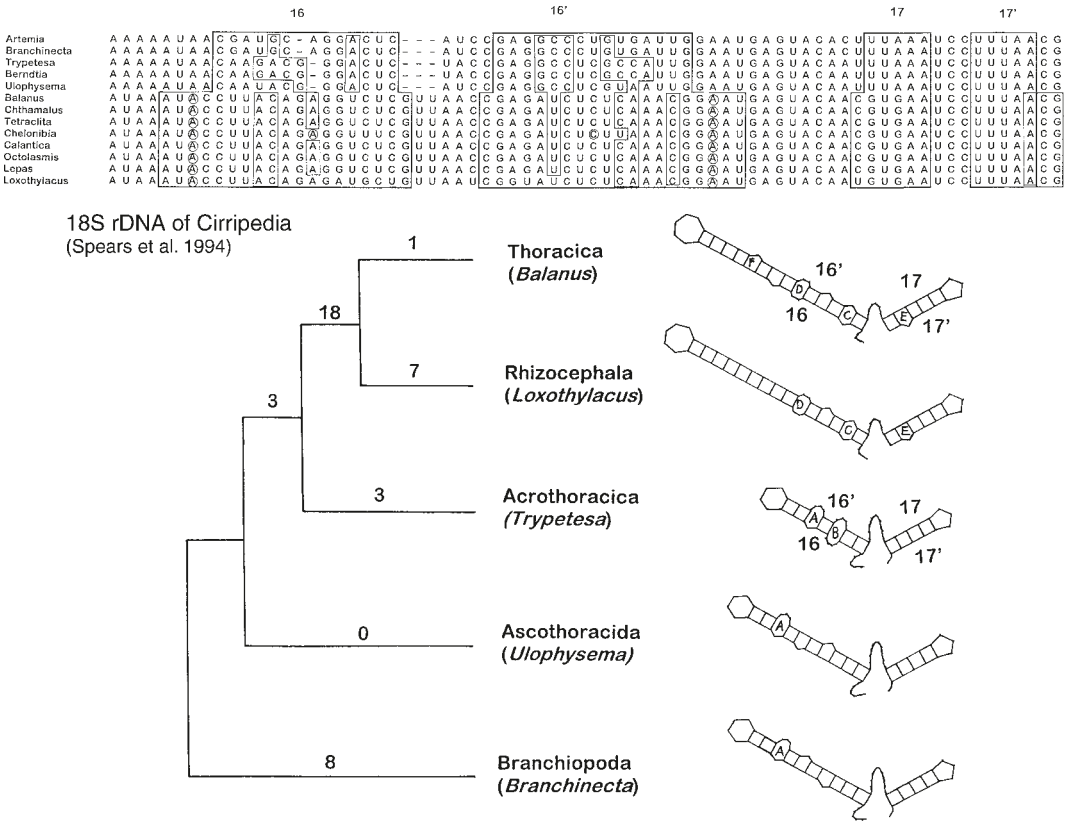


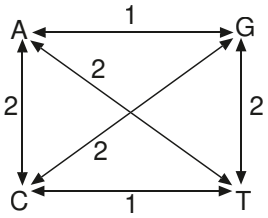
Fig. 140. Insertions in 18S rDNA sequences are unpredictable from the point of view of those who want to model sequence evolution. The example shows the evolution of the V3-region of the sequence of barnacles (Cirripedia: Acrothoracica, Rhizocephala, Thoracica) and the consequences for the secondary structure (bottom; modified after Spears et al. 1994). Numbers on the dendrogram are the number of nucleotide changes (including insertions). Alignment areas shown in frames are complementary and form helices.

methods (ch. 8). In cladistics they did not gain importance. The following methods can be used for *a priori* weighting (Williams 1992): weighting according to the secondary structure and weighting of specific substitutions.

Higher weighting of helical regions

Helical regions of the secondary structure of DNA- or RNA-molecules are thought to evolve under higher selection pressure than regions with unpaired nucleotides. Higher weighting of sequence positions which take part in Watson-Crick base pairing has been justified with this assumption (Wheeler & Honeycutt 1988). In practice, this can be done by counting all paired positions twice or by giving them x-times the weight of unpaired positions. A lower weighting corresponds to the assumption that the substitution of an unpaired base occurs more frequently than that of a paired base (probability of events) or that the probability of homology is x-times higher in paired regions (probability of cognition).

This conception involves three sources of error: weights are chosen arbitrarily and uniformly for all regions of the secondary structure. Whether a weighting scheme simulates the real processes is usually not knowable. It is generally advisable to be cautious: in a comparison of the variability of 18S rRNA molecules of different animal species it can be seen that there exist also variable double-stranded areas as well as conserved single-stranded areas. Obviously, functional constraints do not depend exclusively on base pairing (see Fig.



	A	G	C	T
A	0	1	2	2
G	1	0	2	2
C	2	2	0	1
T	2	2	1	0

Fig. 141. Transformation matrix *(step matrix, cost matrix)* for differential weighting of transitions and transversions.

46). Eventually, the variability of the secondary structure can vary and be larger in some organisms than in others, as known for insertions (see insertions in Fig. 140).

An empirical evidence for the variability of individual positions in a sequence can consist of the number of character state changes per position counted along a reconstructed topology. Topology dependent diagrams of the position specific variability can be calculated with the program MacClade (Maddison & Maddison 1992), for example.

Differential weighting of specific substitutions

A basis of differential weighting is the observation that in nature transitions occur more frequently than transversions, wherefore multiple substitutions and thus analogies as well as the erosion of synapomorphies are to be expected more often with transitions than with transversions (see ch. 2.7.2). This differentiation can be coded with a transformation matrix. When transversions get twice the weight of transitions, the transformation matrix looks like that in Fig. 141. To reconstruct trees using a transformation matrix the algorithms of generalized parsimony are especially suitable (ch. 6.1.2.4).

Weighting of the *probability of events* for **transitions and transversions** means that frequent substitutions get a lower weight because they produce more often chance similarities, they get noisy more rapidly and are not detectable after some time due to multiple hits (erosion of signal). Applying this sort of weighting, it is usually not tested to what extent the erosion of signal occurs. When transitions are found twice as frequently in the alignment, transversions could be weighted twice as high. Doing this it is presupposed

- that the events happened with the same probability for all taxa at all times,
- and that no multiple substitutions mask the real ratios of substitution types (corrections for multiple substitutions are introduced in ch. 8.2.6).

Often different weighting schemes are tested (e.g., for Ts:Tv the ratios 1:2, 1:4, 1:10 are used) in order to select the weights which yield the most plausible dendrogram. It is self explanatory that this method is unsubstantiated and circular (the weighting scheme which supports a preferred hypotheses of phylogeny is favoured).

An extreme form of weighting would be to ignore all transitions. This is equivalent to coding the sequences only for purines and pyrimidines (RY-coding). (Attention: transversion distances may not be the same as distances obtained with an RY alphabet: algorithms implemented in computer programs will possibly differentiate four types of transversions.)

Since information on the real historical substitution processes is generally not available, there is the possibility to weigh positions of an alignment according to whether the visible substitutions are frequent and thus less informative (Schöniger & von Haeseler 1993). A method for weighting of substitution types independent of the alignment position is **combinatorial weighting** (Wheeler 1990; s. appendix 14.2.2). For each pair of nucleotides of a sequence position in two sequences (nucleotide i in sequence 1, nucleotide j in sequence 2), the frequency for $i \rightarrow j$ observed in an alignment is determined and the reciprocal value of the frequency is used as weight for the corresponding type of substitution. One can also count the positions in which the nucleotides i and j occur (**existential weighting** of Williams & Fitch 1990). However, this counting of substitutions in

alignments does not correspond to the number of substitutions occurring in a dendrogram, because analogies in two sequences contribute to the frequency in the same way as synapomorphies. Unrealistic assumptions of differential weighting are listed in appendix 14.2.2.

Similarly, **codon positions** which evolve at different rates can also be evaluated (s. ch. 2.7.2.4). A simple approach consists of a pairwise comparison of sequences to count how often substitutions can be found in the 1st, 2nd, or 3rd codon position of a protein-coding gene. For example, the comparison of two complete mitochondrial genomes of two species of seals (Árnason et al. 1993) has shown that the ratio of substitutions for 1st, 2nd and 3rd codon positions was 2.7:1:16. As for the evaluation of transversions, the applied weights are the reciprocal values (0.37:1:0.06). According to the same principle, a substitution matrix for amino acids found empirically by comparison of proteins can be used for weighting of substitutions (e.g., matrix of Dayhoff (1978): ch. 5.2.2.10). Another method consists of coding in the R-Y-alphabet all codon positions of an alignment which show synonymous substitutions in order to reduce the frequency of chance similarities. The procedure is based on the already mentioned observation that the selection pressure on synonymous substitutions is lower, the corresponding positions evolve more rapidly.

6.3.2 The analogy problem: the creation of polyphyletic groups

The formation of non-monophyletic groups supported mainly by analogies or convergences (ch. 4.2.3) is also a problem in molecular systematics. The analogy problem is called somewhat mystically (though figuratively) "**long branch attraction**" or "the long-branch problem" (Hendy & Penny 1989). "Long branches" or "long edges" are stem lineages in a topology that show a large number of substitutions which cause a replacement of apomorphies (signal erosion!). Also, chance similarities shared by two taxa which are not sister groups accumulate. The source of error relevant for phylogeny inference is the *disproportion* between analogies and homologies: when analogies dominate, false sistergroup relationships supported by noise appear in optimal trees.

The following situations can occur (Felsenstein 1978b, Hendy & Penny 1989):

1) attraction of taxa due to high substitution rates (Fig. 142),
2) attraction of taxa due to shared symplesiomorphies and analogies and also due to lack of competing patterns of apomorphies supporting the correct monophylum (Fig. 143), even when the substitution rates are not different in all lineages,
3) attraction of taxa due to parallel shifts in base frequencies (a special but frequent case of analogies),
4) However (attention!): real sister taxa can indeed share a long stem lineage seducing the observer to the wrong assumption that the grouping has been produced by analogies. This assumption can be tested: a hypothesis of monophyly has to be founded on a number of apomorphic substitutions that is distinctly higher than the background noise (see ch. 6.5, 14.7) or by morphological apomorphies of high quality (ch. 5.1).

The formation of non-monophyletic groups by analogies can easily be illustrated for four taxa and DNA-sequences with a model (Fig. 142). The higher the difference for the probability of substitutions on long and short edges, and the lower the number of distinguishable character states, the more likely is the formation of a non-monophyletic group during tree inference. When the substitution rate q for the short edge supporting a split (middle branch in Fig. 142, lower left) is plotted against the rate p of the neighbouring long edges, k being the number of character states (in case of DNA: k=4), then the graph

$$q = p^2/(k-1)$$

describes the area in which groupings occur due to analogies (formula by Mishler 1994). This area is also called the ***Felsenstein-zone*** after the discoverer (Felsenstein 1978b).

Long branches can also fuse to a false stem lineage when the apparent sister monophylum is also separated by a long branch (Fig. 143). In this case on the lineage to {C,D} the synapomorphies that would help to recover the clade {B,C,D} eroded and therefore ((A,B),(C,D)) is the most parsimonious solution, even though {C,D} is supported only by plesiomorphies.

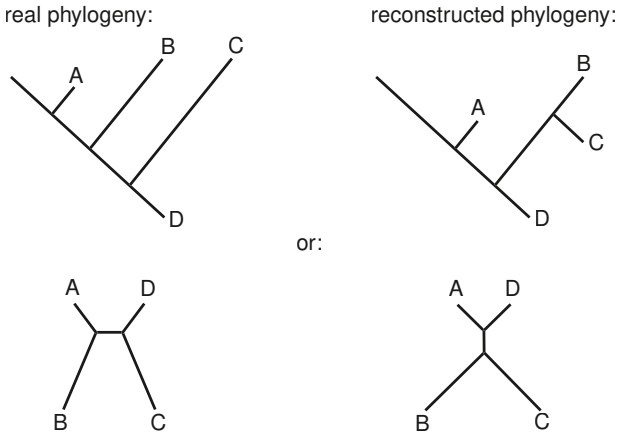


Fig. 142. Effect of analogies that emerge in lineages with high substitution rates: the length of the branches symbolizes the real number of substitutions. The analogies support false groupings in reconstructed phylogenies. A-D: recent taxa.

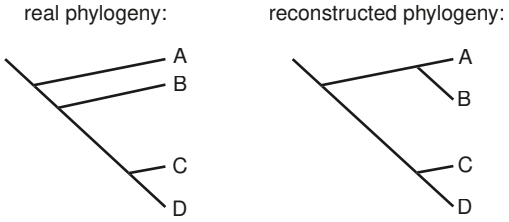


Fig. 143. A-D are recent species, branch lengths symbolize the divergence time. The wrong grouping {A, B} in the reconstructed phylogeny is produced by analogies and/or symplesiomorphies while synapomorphies shared by {B, C, D} eroded on the stem lineage of {C, D} (see also Fig. 145).

An erroneous grouping of taxa can also occur when real monophyla are not "neighbours" as in Fig. 143. Examples: a 18S rDNA analysis of Metazoa yielded a sistergroup relationship between Nematoda and related forms (Cycloneuralia) and Arthropoda ("Ecdysozoa-hypothesis"). Both are taxa with "long" stem lineages. The authors (Aguinaldo et al. 1997) thought that the presence of a cuticle is a further homology shared by both taxa, but ignored the large number of potential synapomorphies occurring in Annelida and Arthropoda. A new analysis of the molecular dataset has shown that the alignment is very noisy and probably only analogies support the clade Ecdysozoa (Wägele et al. 1999). – Many evolutionary novelties like a complex tooth structure, a spiralled cochlea, a corpus callosum in the brain, vivipary, milk glands with nipples (see Thenius 1979, Cifelli 1993) present in the Theria (Marsupialia + Placentalia) prove that the grouping Monotremata + Marsupialia (= Marsupionta; Fig. 144) found in molecular systematic analyses (Janke et al. 1996) most likely is not monophyletic. The stunning number of probably homologous novelties in the soft anatomy of the Theria cannot be overlooked and makes the result of the molecular systematic study implausible. The molecular support of the Marsupionta (analyses of mitochondrial genomes) is partly based on base composition bias (D. Penny, pers. comm.). Analyses of other genes confirm the traditional classification (monophyly of Theria: Killian et al. 2001).

It has been noted that errors occur when real sistergroups have long branches. In maximum likelihood analyses such a sistergroup-relationship may not be recovered (Siddall 1998, Pol & Siddall 2001: "long-branch repulsion"), while with maximum parsimony the taxa are correctly grouped together. The cause of this effect is not known.

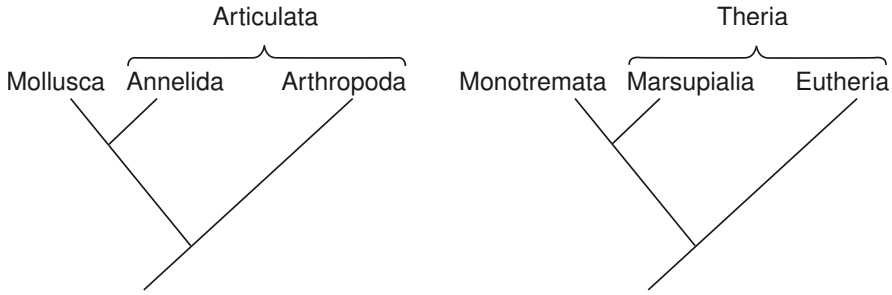


Fig. 144. Dendrograms estimated from DNA sequences with typical groupings that can be caused by symplesiomorphies or chance similarities. For the Articulata and Theria, clades that are incompatible with some published gene trees, many morphological apomorphies are known. Arthropods are very diverse and highly evolved, the same is true for the Eutheria, while the basal diversification in these trees occurred probably in comparatively short time. The historical situation is similar to that seen in Fig. 143.

Elimination of mistakes: a spectral analysis can clarify whether the signal in favour of a group is markedly higher than the background noise which supports other groupings with chance similarities (ch. 6.5). Often also a glance at the alignment is sufficient to recognize the high variability of positions. According to our experiences, spectra estimated for alignments which have many multiple substitutions do not show clear signals (Wägele & Rödding 1998). Furthermore one can test whether a specific sequence shares in spectra similarities with various groups of sequences. A sequence that appears in many different and incompatible splits has high substitution rates and shows many chance similarities with unrelated groups of species. Finally, also the "relative rate test" (ch. 14.8) gives hints for the presence of "long branches", but only if variable positions are not saturated with substitutions. When the suspicion arises that "long branches" are present, one can either eliminate the taxa in question and search for species that represent a "slower" clade or one has to sequence a different gene that evolved with fewer substitutions.

6.3.3 The symplesiomorphy trap: paraphyletic groups

The support of groups by plesiomorphic character states can be a cause for the postulation of implausible or inconsistent hypotheses. This is the reason why Hennig stressed the difference between apomorphies and plesiomorphies. In molecular systematics the effect of plesiomorphies has been largely ignored. Several examples exist in the published literature where the source of error cannot be detected without careful search for erosion of apomorphies (a long-branch-effect), for example. The incongruence could have been discovered with additional information such as data from anatomy or from the fossil record.

If the same data (with the same bias) are used for analyses with different reconstruction methods (e.g., NJ-, MP-, ML-methods), the implausible dendrogram will be obtained repeatedly. It even might be supported by high bootstrap-values despite its contradiction with the real phylogeny. A cause for the consistently wrong results can be the presence of characters which have the distribution of a synapomorphy for a wrong (in reality non-monophyletic) grouping. Symplesiomorphies can play the role of fake synapomorphies. Fig. 145 explains how this is possible.

Shared old characters of the common ancestor of taxa A-E and still present in taxa A and B in Fig. 145 evolved further in the stem lineage of {C, D, E}, so that three character states can be distinguished *in relation to the monophylum {C, D, E}*: distant outgroup characters (state 0, not shown) plesiomorphies (state 1), and apomorphies (state 2). The plesiomorphies cannot be identified to be apomorphies of {A, B, C, D, E} and they support the wrong group (a paraphylum). This occurs when

– plesiomorphies do not occur in the more distant outgroups, and apomorphies of the real sistergroup-relationship B + {C, D, E} (Fig. 145) are not present or too rare in relation to the number of symplesiomorphies. This happens

when the stem lineage of B + {C, D, E} is short or when the substitution rate is very low; or
– one of two sister taxa (stemline of {C, D, E} in comparison with B) shows many more character state changes, for example, after an intensive phase of adaptive radiation, with the effect that many of the older characters are not conserved any more (erosion of apomorphies; e.g., Arthropoda in comparison to Annelida).

For the specific situation in Fig. 145, parsimony analyses would give the same tree lengths (6 steps) for both topologies. When additionally only one analogy of the species A and B occurs, the wrong phylogenetic tree is the most parsimonious one. As analogies are to be expected regularly in DNA-sequences, the symplesiomorphy trap is effective.

Correction of mistakes: The effect of symplesiomorphies can be a long-branch attraction similar to the one caused by accumulating analogies. However, as plesiomorphies are homologies and do not have the characteristics of chance similarities (i.e. slow accumulation in all possible splits of a dataset), there can appear a signal in spectral analyses supporting the paraphyletic group which is clearly above the level of "background noise". For this reason each method of tree construction will find the apparently monophyletic group. Long internal branches can be shortened by addition of *further taxa* which introduce more symplesiomorphies into the dataset. These can be closely related outgroup taxa (X in Fig. 145) or ingroup taxa (Y in Fig. 145). The presence of symplesiomorphies in further taxa reduces the number of supporting characters for the paraphylum. An example is illustrated in Fig. 146.

More comprehensive taxon sampling often has an effect on tree topologies (Fig. 147). Unfortunately, better taxon sampling is not possible in molecular systematics when stem lineage representatives are extinct. Incongruence between molecular trees and topologies estimated from morphological and paleontological data should be the motive to search for symplesiomorphy effects.

These observations explain the sources of mistakes which are to be expected when a topology contains very long branches: plesiomorphies or

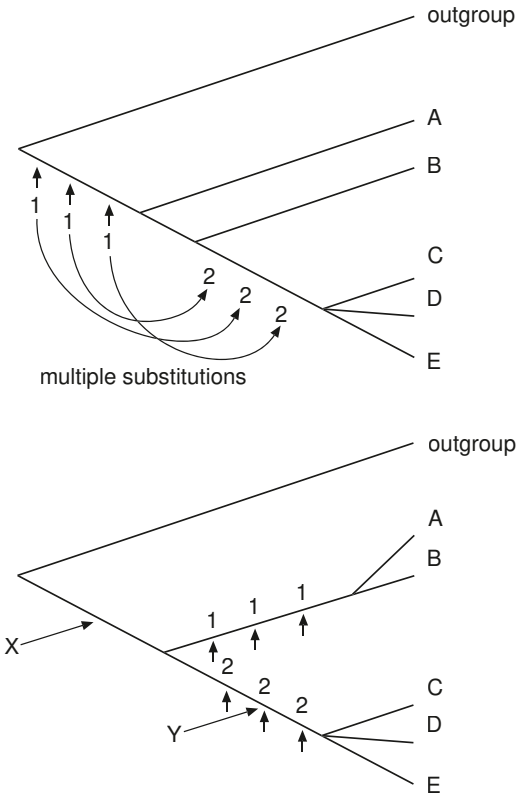


Fig. 145. Situations in which the real phylogeny (top) cannot be reconstructed: the wrong grouping {A, B} (lower topology) is supported by symplesiomorphies (characters of type 1), which are not detected because the informative characters were substituted by new ones (characters of type 2). The effect of symplesiomorphies can be compensated by addition of further taxa (X and/or Y).

analogies may be misleading and are difficult to detect. To avoid artifacts it is important to find out whether some taxa evolved rapidly. This can be tested with the relative rate test and related methods (ch. 14.8). However, by comparing branches with the same locally variable positions, this test may not discover a difference in the number of multiple substitutions (i.e., when the same variable positions were hit with different frequencies).

6.3.4. Using alignment gaps

Alignment gaps are the result of insertions or deletions which are only present in some of the

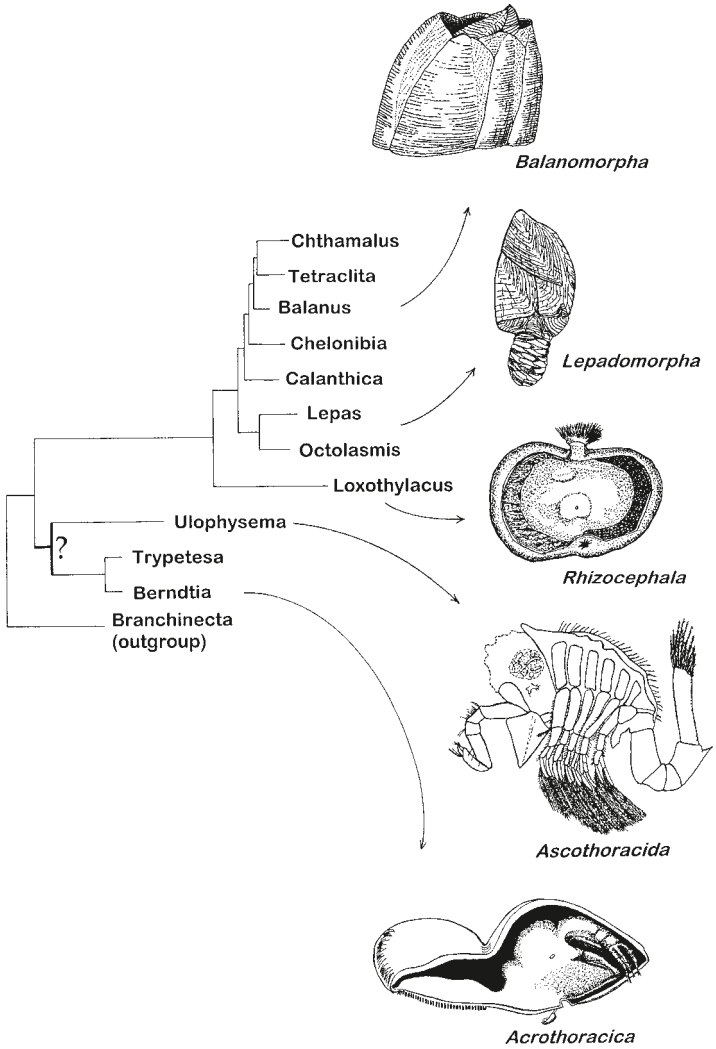


Fig. 146. Phylogeny of Cirripedia (barnacles) and Acrothoracica reconstructed from 18S rDNA sequences (after Spears et al. 1994). The grouping {Ascothoracida, Acrothoracica} seen in this gene tree is not plausible, because the Ascothoracida have the most primitive morphology of all species considered and the Acrothoracica share derived characters with higher derived species of this topology. They show the same adaptations to a sessile mode of life as the other Cirripedia. Actually, the topology changes and the Ascothoracida appear basally as sistergroup to {Ascothoracida, Rhizocephala, Lepadomorpha, Balanomorpha} when more outgroups are considered (Wägele 1996).

sequences of an alignment (s. Fig. 148). As gaps* often occur in regions of uncertain positional homology, it has frequently been recommended to exclude such sequence areas for phylogenetic analyses. However, in some cases these areas are especially informative, for example when individual clades can be distinguished due to the presence of certain insertions or deletions (e.g., in rDNA-sequences: Fig. 140). Another argument raised is that gaps bear no empirical evidence

* Since in positions with gaps it is often not clear whether a deletion in some species or an insertion in the other ones produced the pattern, the pattern is also called an "indel".

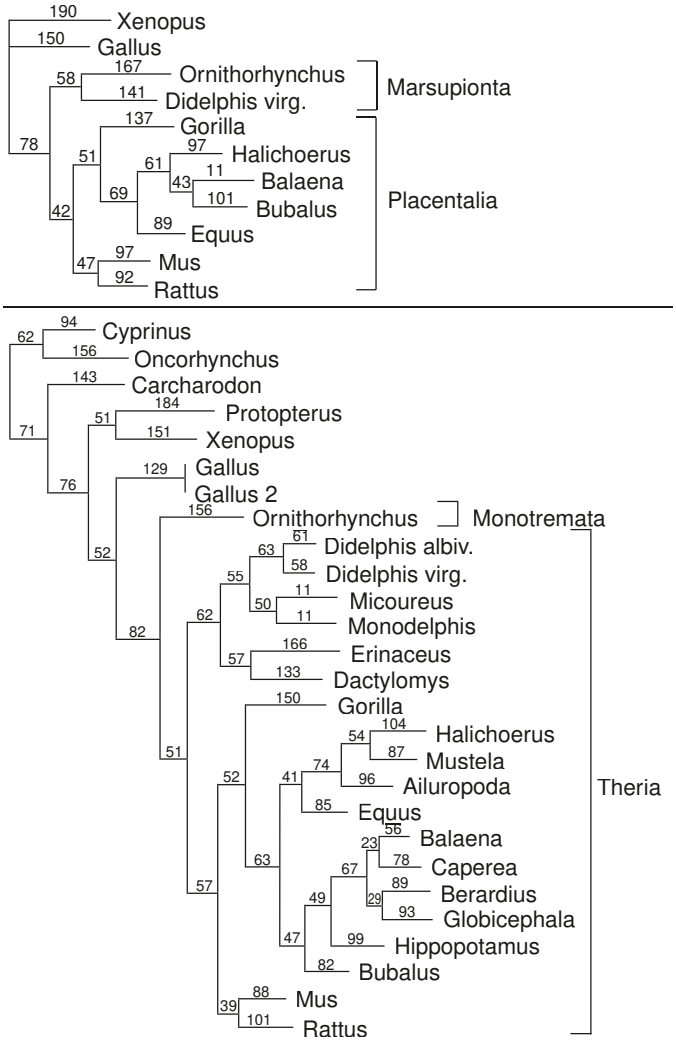


Fig. 147. Demonstration of the influence of taxon sampling on a topology of cytochrome b sequences of mammals (most parsimonious MP-topology with the number of character changes shown on branches). The topology at the top supports a sistergroup-relationship Marsupionta/Placentalia, the lower one a relationship Monotremata/Theria. The alignment is not very informative, note the implausible arrangement of some taxa of the Theria.

(absence is not a character) and therefore have to be excluded. However, gaps are often the counterpart of informative insertions.

Depending on the structure of the dataset, the topologies obtained with different gap treatments can differ markedly (Fig. 148). Be aware of the implied hypotheses: each way of gap coding represents some hypothesis about the probability that characters are homologies. The systematist has to decide if there is some evidence

- for an insertion being the result of a single event (as in the case of translocations). Check
- if the elongation of an insertion or gap could be the result of stepwise evolution that can be coded as multistate character,
- if indels may occur convergently in many lineages,
- and if the alignment is reliable for the region that contains gaps.

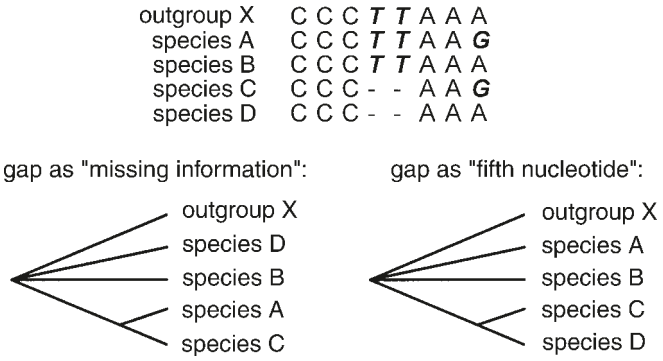


Fig. 148. Model for an alignment illustrating the effects of deletions and insertions in parsimony methods (see also Fig. 167). The species X is defined as the outgroup. Species C and D have a common ancestor with a loss mutation (deletion of ***TT***), species A and C show a convergence (***G***). With the parsimony method the correct group {C, D} is found when gaps are coded as homologies ("fifth nucleotide"), because the number of potential synapomorphies (loss of two "***T***") is larger than in the group {A, C} (mutation A→G). However, when the gaps are coded as "missing information", the relevant positions do not have an effect and then the sequences A and C are more similar.

Therefore, there are two different levels where mistakes can occur:

– *positional homology* may be incorrect,
– *character state homology* may be uncertain.

Any further discussion about the use of indel characters is only meaningful if we assume that positional homology has been determined correctly with high probability. The lower this probability is, the lower should be the weight for any character transformation, and in many cases it is wise to exclude the alignment regions with variable sequence length.

Popular parsimony programs (e.g., PAUP, Swofford 1990) allow the user to decide whether gaps are treated as "missing information" or as "**fifth nucleotide**". In the first case, tree inference is more strongly influenced by positions without gaps. This option is to be chosen when alignment regions seem to be ambiguous, when it is possible to align gaps differently even with constant optimality criteria, or in short, when the probability of homology is low for gaps. However, whenever it is probable that a homologous insertion or deletion is present (this is the case when longer, *conserved* sequence sections are affected), the single gap can be coded as a discrete character ("fifth nucleotide"). If possible, a step-matrix should be used to give the gain of a nucleotide a higher weight than the loss to reflect the differences in probability of homology (see below).

An objection against the use of positions with gaps is often raised because a single event (insertion or deletion) can affect several positions (for example: −−−−− → AAGAT). Systematists who want to evaluate events separately would count such an insertion as a single character. To do this, each insertion can be recoded as a single character (other variants are discussed in Young & Healy 2003). Whereas if the probability of homology is considered phenomenologically, the specific pattern "AAGAT" has to be weighted higher than the single character "A" (see ch. 5.1). Therefore it is recommended to count positions with insertions individually for MP-analysis. This requires differential character weighting: it has to be taken into account that there exists an asymmetry in the estimation of homology of character states. In contrast to insertions, deletions produce an unspecific pattern (e.g., AAGAT → −−−−−), which offers no details for the comparison of alternative hypotheses of homology. Deletions should get a lower weight than insertions in a phenomenological analysis (e.g., 1/n instead of 1, n being the number of alternative character states).

Another problem is that areas with ambiguous alignment may show for some closely related species clearly homologous patterns (e.g., exactly the same insertion), while the corresponding character states are uncertain for other species of the alignment:

```
Species_1    GG---TAGCT--CC
Species_2    GG--AATG-T--CC
Species_3    GGTGAGACCTTACC
Species_4    GGTGAGACCTTACC
Species_5    GG--AGTC----CC
Species_6    GG--TGTGC---CC
```

Deleting these positions from the alignment would mean a loss of information. A solution is a recoding of these positions:

```
Species_1    GG??????????CC
Species_2    GG??????????CC
Species_3    GGTGAGACCTTACC
Species_4    GGTGAGACCTTACC
Species_5    GG??????????CC
Species_6    GG??????????CC
```

This part of the alignment will support a partition separating the group {species_3, species_4}. Such clade-specific signatures can be extracted from an alignment to add them at the end of the alignment, then the corresponding ambiguous region is deleted.

Another proposal is to recode ambiguous regions each as single character with as many states as there are specific sequences in the region (Lutzoni et al. 2000). For example, all sequences of the type AAGGTT would be coded with state 1, all sequences AAGAT with state 2, etc. Then a step matrix is constructed to weigh each change of character state. There may be many ways to weigh these steps in relation to sequence similarity and estimated number of changes. Since nothing is known about the probability that indels evolve, models are not useful and only a phenomenological analysis is possible.

The consideration of alignment gaps in distance methods is discussed in ch. 8.2.4. In contrast to parsimony methods, model-dependent distance or maximum likelihood methods need specific models for the evolution of insertions or deletions. This is the reason why gap positions are usually ignored for these methods.

When using tree constructing software one should be informed about the way gaps are treated. With the optimization alignment program POY, for example, indel changes may be weighted more heavily than substitutions (depending on the selected step matrix), while MALIGN uses gaps as fifth character state (Young & Healy 2003).

```
Trichinella     -CGTTTTA-- -----GCTT- CTGTCGCCT- -TTCGCCCAA TT--CGC-TA
Gordius         TAAACTTA-- -G---C-TT- TTGTCCCC-- CACGGCCCAA TT--CGC-AA
Priapulus       -GGGATAT-- -----C-CT- CTGTTCAC-T ATCCGCTAAT TG--CGC-AA
Pycnophyes      -CTCCGGTGC -----C-CT- TTGTTGCC-- ATCCGCTAAC TG--CACG-A
Euperipatoides  -CCTAATT-G G-GTGGCATC TTGTCCCCTG CGGCGCTAGT CGGTCGC-GA
Macrobiotus     -TATAA-T-G G----C-AT- TTGTTTCCT- TACCGCCTGT ---CCGC-AC
Aphonopelma     CTTCCTTA-G -----C-GG- TTGTCCCC-- TACCGCCCA- -G--TGCGTG
Scolopendra     -TGCATTA-- -----C--TC TTGTCCCC-- CGTCGCCCAC CGAGCGCG?G
Tenebrio        TAGAATTA-G G----GTCT- CTGTCTCC-- CGTCGCTCGC -TGTCGT-CG
Panulirus       -GGACTT--C C------CT- CTGTCTCC-- -GCCGCTAA- -TCGCGTT-G
Brachionus      -TCTTAGTA- -------TC- ACAGTTATCA TATTATTAGT AG--TAT--G
Liolophura      -GTTCA-T-- -C-----TC- CCCGTTAT-- TGCTATTGAC ATC-TAT-GG
Placopecten     CTTTTT-T-- -----G-TCA TCCGTTAT-- TGTTATTGGC -TC-TAT-CG
Enchytraeus     -GTTTA-T-- C-CGG-GAC- ACAGTGAT-- GGTTATTGGT ATT-TAT-CG
Stylaria        -ATTAA-T-- -------AG- ACCGTTAC-- TGCTATTGGT -TT-TAT-GG
Glycera         -TTTTA-T-A ------GGC- ACCGTTAT-- TGTTATTGAT -TC-TAT-TG
Terebratalia    -TTTCA-T-- ----C-GGC- CCCGTTAT-- TGCTATTGA- -TCATATT-G
Antedon         -ATGTAC--- ----CGG-T- CCCTCCTT-- TGTTGCTAG- -TC-TATT-A
Tripedalia      -ACTT-GT-- -----G--G- TCATGTCT-- T-CTATTAGT AT--TAT--C
Stenostomum     -GTTTGGC-- -----GAAC- TCCGTGCC-- GTCTATTGGG AC--TAC-GG
```

Fig. 149. Sequence positions with nucleotides identified with parsimony analysis as putative apomorphies of the clade Ecdysozoa (18SrDNA alignment of Aguinaldo et al. 1997). Genus names above the horizontal line are representatives of the presumed monophylum Ecdysozoa. Note that even though these positions change character states on the branch leading to the Ecdysozoa in the most parsimonious tree, most do not really fit to the split between ingroup and outgroup, the columns are very noisy. A single binary position fitting to the split is highlighted. To find out if this pattern is signal or noise spectral analysis is the best tool (ch. 6.5).

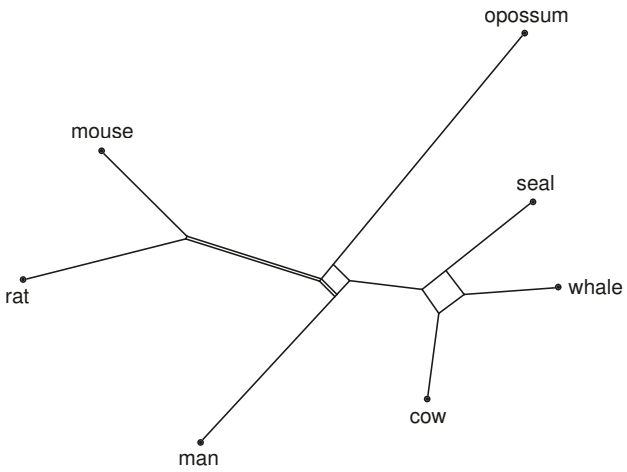


Fig. 150. Example of a split-graph. Diagram for the mitochondrial ND2 gene of some mammals (modified after Wetzel 1995).

6.3.5 Potential apomorphies

Apomorphies occur in DNA sequences as well as in morphological datasets. Lists of potential apomorphies for all branches of a rooted topology can be compiled with suitable computer programs (e.g., with PAUP). It is strongly recommended to select and inspect the relevant sequence positions for groupings which do not seem to be plausible considering other information (see "plausibility of hypotheses" in ch. 10). A matrix with a pattern of putative apomorphies can show how many perfectly fitting, i.e. binary positions match the considered split, and how many positions do not appear to support the split clearly because they are too variable (Fig. 149). The number of potential apomorphies (which also corresponds to the "branch length" in the parsimony method) is often higher than the number of positions which unequivocally fit to a split, because in parsimony analyses every character state change that is constructed with this method is counted independently of character quality. Branch length depends, e.g., on sequence length and on the species composition of the dataset. Whether the pattern of positions supporting an implausible clade could be caused by chance similarities or by a distinct homology signal can be tested with spectral analysis (ch. 6.5).

6.3.6 Lake's method

Lake (1987) suggested a method of sequence analysis ("evolutionary parsimony"), which is based on the comparison of quartets of sequences, whereby only transversions are considered. It requires the (unrealistic) assumption that transversion rates for different nucleotides are equal. As there are only 3 alternative topologies for 4 taxa, it can be tested for each combination of 4 sequences which of the 3 possible topologies has the best support. The method is not used very much (see also Felsenstein 1991, Swofford et al. 1996). The known inefficiency of the method is based on the fact that it uses only part of the information of a dataset and therefore requires larger alignments than other methods.

6.4 Split-decomposition

The following method is explained in this larger chapter (phenomenological methods) because it can be used for exploratory analyses of morphological and other types data. However, it is also possible to construct graphs containing edge lengths estimated with model-dependent methods.

Split decomposition (Bandelt & Dress 1992) permits the inclusion of alternative incompatible topologies in a single graph. In a dataset with n terminal taxa at most $2^{n-1}-1$ splits can occur, however, in practice far less splits are really represented by character states. When there exists only exactly one dichotomous topology for a dataset, the number of splits is equal to the number of branches (2n–3; ch. 3.4). When analogies occur, several alternative dichotomous topologies are supported by the dataset. The number of topologies represented in real data can be very different. Split-decomposition visualizes the conflict within an alignment and can be used to compare conflicting evidence within different datasets and is therefore more general than the MP-method. In the following, **d-splits** are explained (for "parsimony splits" see Bandelt & Dress 1993), a more detailed description can be found in the appendix (ch. 14.4).

A basis for the construction of d-split diagrams are measures for the distance between pairs of terminal taxa. Discrete characters (homologies evaluated phenomenologically) can be used as well if differences are coded like genetic distances. The distance can be:

- the number of visible sequence differences (Hamming-distance, see ch. 14.3.1),
- the estimated number of substitution events (evolutionary distance, see models of sequence evolution, ch. 8.1, 14.1),
- the number of split-supporting sequence positions,
- the number of morphological character changes.

The user of these methods should be aware of the specific assumptions implied by different distance measures (see assumptions required by models of character evolution: Fig. 159).

The split-graph is constructed from distance data. As the topology of a dichotomous unrooted tree is already defined by the relationship between groups of 4 terminal taxa (ch. 14.3.3), the analysis can be performed with quartets. For each group of four terminal taxa i, j, k, l it is tested which of the three possible splits ({(i,j),(k,l)} or {(i,l),(k,j)} or {(i,k),(j,l)}) has the weakest support. Checking quartets, of the three possible splits the two best supported ones are retained, the third split is not considered (ch. 14.4). In practice it is not necessary to test all theoretically possible groupings. It is sufficient to evaluate all splits present in a given dataset and to combine the compatible and weakly compatible ones in one graph (see also Fig. 105, Fig. 195)

Fig. 150 shows that the genetic distance between the opossum on the one hand and the Eutheria (all other taxa of Fig. 150) on the other hand is large in this alignment of ND2 sequences. We can state that this split is well supported. The relationship between ungulates and the Carnivora, however, cannot be settled because there are supporting characters for the split {(seal, whale), (cow, other mammals)} as well as for the split {(cow, whale), (seal, other mammals)}. Each of these splits is represented by two parallel edges, together they can be depicted as a rectangle. A similar conflict exists for the closest neighbour of rodents, where we can choose between *Homo* and opossum. The correct interpretation therefore is: this dataset contains too many contradicting characters to allow the selection of well supported hypotheses for some of the relationships using the ND2 gene and distance measures.

In the original concept of split-decomposition (Bandelt & Dress 1992), binary characters have the main effect on the topology of the graph (see Fig. 197). Slightly noisy characters (sequence positions with more than 2 states) are not considered, wherefore in many cases only "star diagrams" are obtained. The method proved to be useful in combination with spectral analysis (ch. 6.5).

An older method designed to find patterns of compatible splits is the clique-method (ch. 14.5). Prerequisite is a binary coding of characters. It has the disadvantage of considering only those characters of a dataset which support a majority of mutually compatible splits. All other characters are not considered for the reconstruction of dendrograms, conflicting information is lost. Networks are not constructed. Models of sequence evolution are not used, in contrast to the spectral analysis discussed in ch. 6.5. Due to the necessity to use binary characters, noisy positions cannot be considered or they must be coded in the RY-alphabet.

6.5 Spectra

6.5.1 Basics

Spectra are useful to explore the information content of data without reference to a known phylogeny. In phylogenetics, a spectrum (Figs. 153, 154, 170) is a graphic representation of data obtained for bipartitions of all terminal taxa. For example, for each bipartition visible or estimated distances between the two groups of taxa are shown, with or without corrections estimated from substitution probabilities, or the number of supporting positions for all splits of a dataset is represented. A dendrogram is not required for this analysis.

In the following we consider the character states in positions phenomenologically (compare Wägele & Rödding 1998). Splits are formed by real apomorphies, when we are dealing with bipartitions each separating one real monophylum from the remaining terminal taxa. Additionally, real data contain chance similarities, convergences, and symplesiomorphies that in most cases support non-monophyletic groups and which in phylogenetic systematics cannot be used to substantiate a hypothesis of monophyly. Because we do not want to refer to a known topology, all informative characters favouring a split are treated equally and are called together the "**number of supporting positions**". We want to find, for example, in a given alignment all splits that have a high number of supporting positions.

Considering a topology of four species and the possible distribution of the character states for a binary character, there are 16 possible character combinations:

species	patterns of character states															
1	0	1	0	1	0	1	1	0	1	0	1	0	1	0	0	1
2	0	1	1	0	1	0	1	0	0	1	0	1	1	0	0	1
3	1	0	0	1	1	0	1	0	0	1	1	0	1	0	1	0
4	1	0	1	0	0	1	1	0	0	1	1	0	0	1	0	1
	s	*s*	*i*	*i*	*i*	*i*	*c*	*c*	*t*	*t*	*t*	*t*	*t*	*t*	*t*	*t*

Fig. 151. Table with all possible combinations of character states for a binary character and 4 species. For the split {(1,2),(3,4)} there are 2 supporting patterns (*s*) and 4 incompatible patterns (*i*), 8 patterns are trivial (*t*), 2 are conserved (*c*). The 16 patterns correspond to 8 possible splits. The splits *s* and *i* are non-trivial splits.

It can be seen in this table that considering four species in a given dichotomous topology (e.g., split S = {(1,2),(3,4)}), there exist for this split two incompatible topologies (split {(1,4),(2,3)} and split {(1,3),(2,4)}), each supported by the same number of possible patterns. When all patterns occur with the same frequency, then there are two supporting (*s*) and four incompatible (*i*) patterns for each split. The trivial patterns (*t*) support no groups, but separate only individual species. They would determine the length of terminal branches. The conserved patterns (*c*) do not contribute to the shape of a topology.

It is the aim of spectral analyses to identify those patterns and relevant sequence positions which unequivocally support a split in a given dataset. Starting from the consideration explained above, one could estimate for a given number of species the probability for the occurrence of patterns in a single sequence position which unequivocally support a specific split (in the table of Fig. 151 these are 2 of 16 possible patterns for the split {(1,2),(3,4)}). In practice, however, one proceeds the other way round and searches for splits represented by real patterns in a given alignment. The analysis of spectra can be done phenomenologically (ch. 6.5.2) or using substitution models (ch. 8.5). The phenomenological analysis is an attempt to imitate the type of character analysis used by morphologists, namely to evaluate characters on the basis of their complexity. Complexity becomes visible in alignments when the supporting positions of a split are selected and grouped to a separate pattern (see below).

6.5.2 Analysis of spectra of supporting positions

This method serves the estimation of the information content of aligned DNA sequences. It is a phenomenological analysis which aims to find the number of supporting positions for each split contained in a dataset if noise is taken into account. A difference from tree construction methods is that it is possible to visualize how many *incompatible* splits are contained in a dataset, so that one can see whether a putative monophylum is clearly better supported than random combinations of terminal taxa, i.e. if support is beyond

the background noise. Unlike d-split decomposition (ch. 6.4, appendix 14.4), the supporting positions do not have to be binary (only 2 character states). Fig. 152 shows which alternative patterns of nucleotides could occur in single sequence positions.

It is obvious that **conserved** (invariable) positions do not contain phylogenetic information, in **very noisy** positions the information is not recognizable, while **symmetric** positions conserve a plesiomorphic and an apomorphic character state (as long as this pattern is not a product of chance). **Asymmetric** positions result when the plesiomorphy of the outgroup taxa is noisy (due to substitutions occurring after separation from the last common ancestor of all outgroup taxa), while the apomorphy of the ingroup is conserved.

For an analysis of the phylogenetic signal conserved in an alignment the following steps are performed:

- Search in each position of an alignment X the groups of taxa which share the same nucleotide. Define alternatively each of these groups as potential ingroup *A*. Define the split $S = \{A,B\}$ with $A \cup B = X$ and group *B* as potential outgroup.
- Name those positions which cause the split the "supporting positions" of the split or the "potential apomorphies" of the putative ingroup.
- Include supporting positions which contain in ingroup sequences some deviations from the consensus character state of the ingroup. These are noisy positions with potential autapomorphies in single ingroup sequences, the ingroup consensus character state is the potential state of the ingroup ground pattern. Allow also in supporting positions the occurrence of single convergences/analogies in sequences of the outgroup to the ingroup consensus state (i.e. homoplasies in single outgroup sequences). Both types of deviations are defined as "noise".
- Positions which show an accumulation of potential symplesiomorphies (outgroup character states) in single sequences of the ingroup cannot be considered to be supporting positions. Such an accumulation indicates that the corresponding sequence belongs to the outgroup or that the sequence conserves a

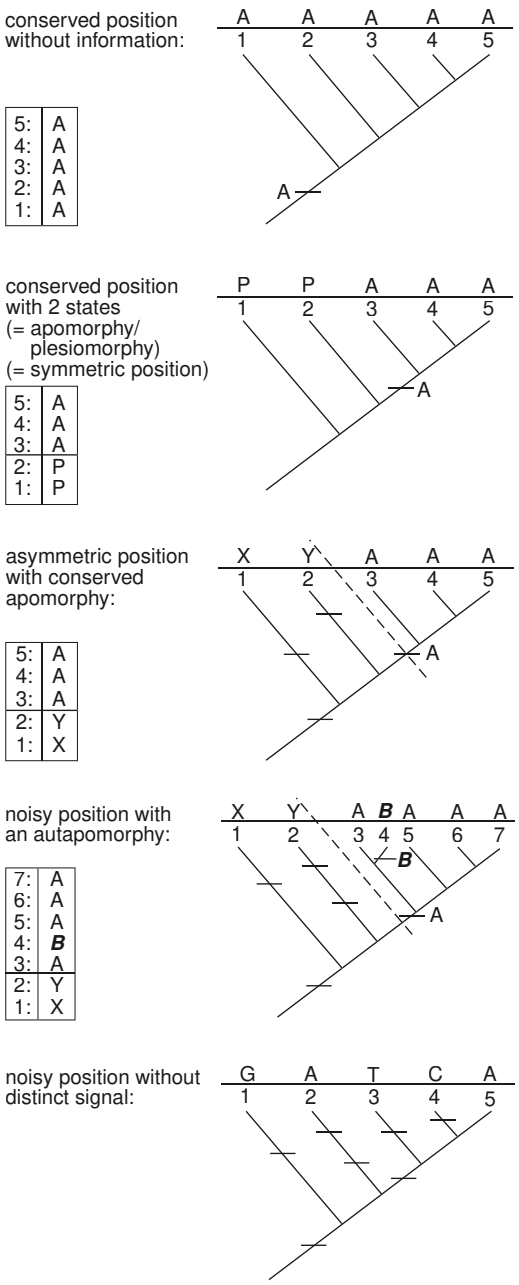


Fig. 152. Patterns in single sequence positions of an alignment of DNA sequences resulting from a series of evolutionary processes. A, B, P, X and Y are character states. A: putative apomorphies; P: putative plesiomorphies; X, Y: states of unknown phylogenetic value (noise); G, A, T, C in the last graph: nucleotides.

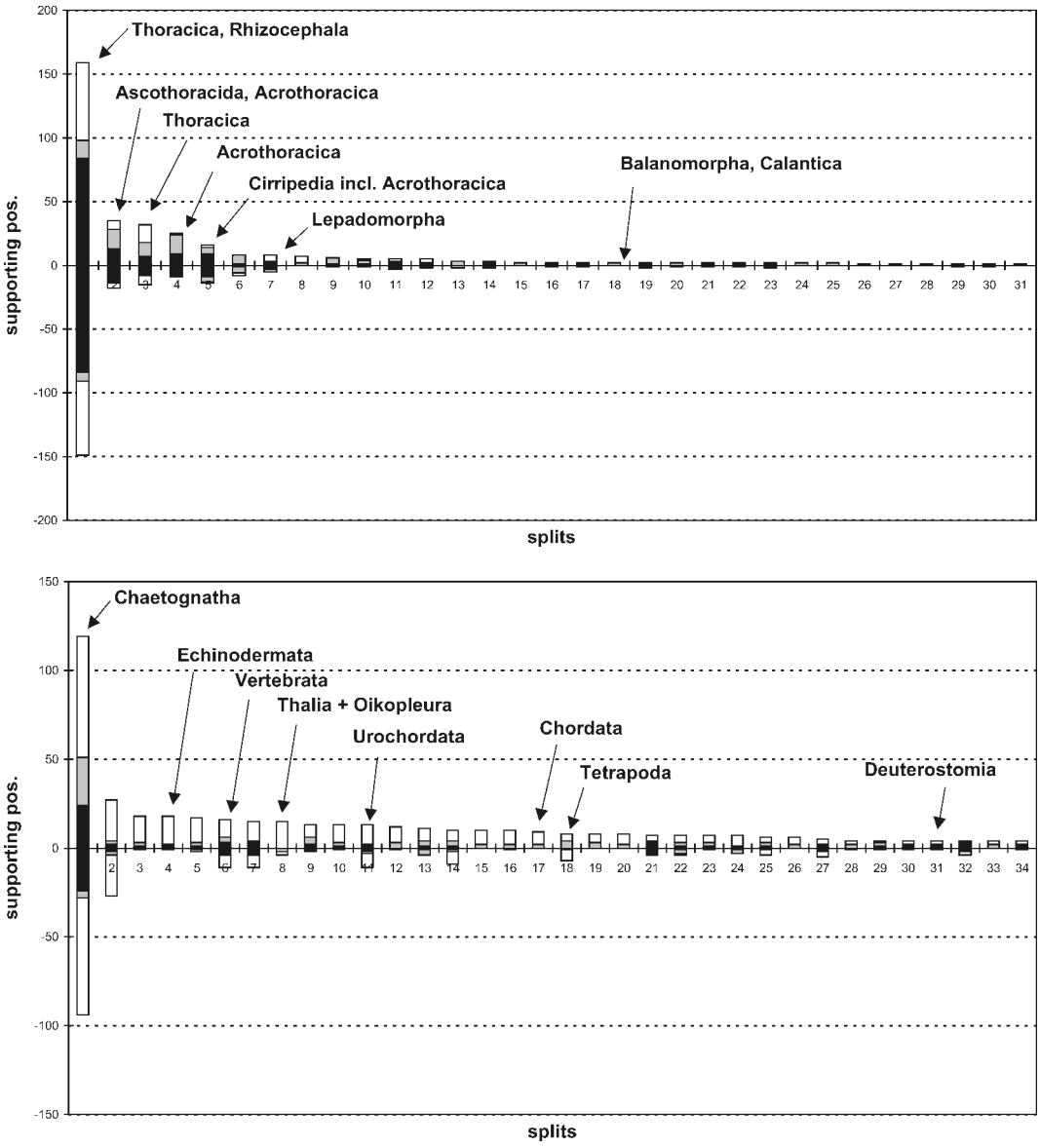


Fig. 153. Spectra of supporting positions: in the upper diagram most of the mutually compatible groups (marked with arrows) which also occur in reconstructed dendrograms are at the same time the splits with the best support. In contrast, support for most of the named splits of the lower diagram does not differ from the support for random combinations of terminal taxa. The correspondent alignment is not suitable for phylogeny inference (from Wägele & Rödding 1998, original data from Spears et al. 1994 and Wada & Satoh 1994). In these diagrams, column height indicates the number of supporting splits, splits are ranked by quality of signal-like patterns. Each split consists of two groups, the support for the weaker group is indicated below the x-axis.

larger number of plesiomorphies.

– Limit the number of deviations in such a way that with high probability they form patterns that can be explained by noise (patterns composed by randomly distributed similarities, Fig. 115). Deviations should not form non-random (signal-like) patterns within the pattern of supporting positions.

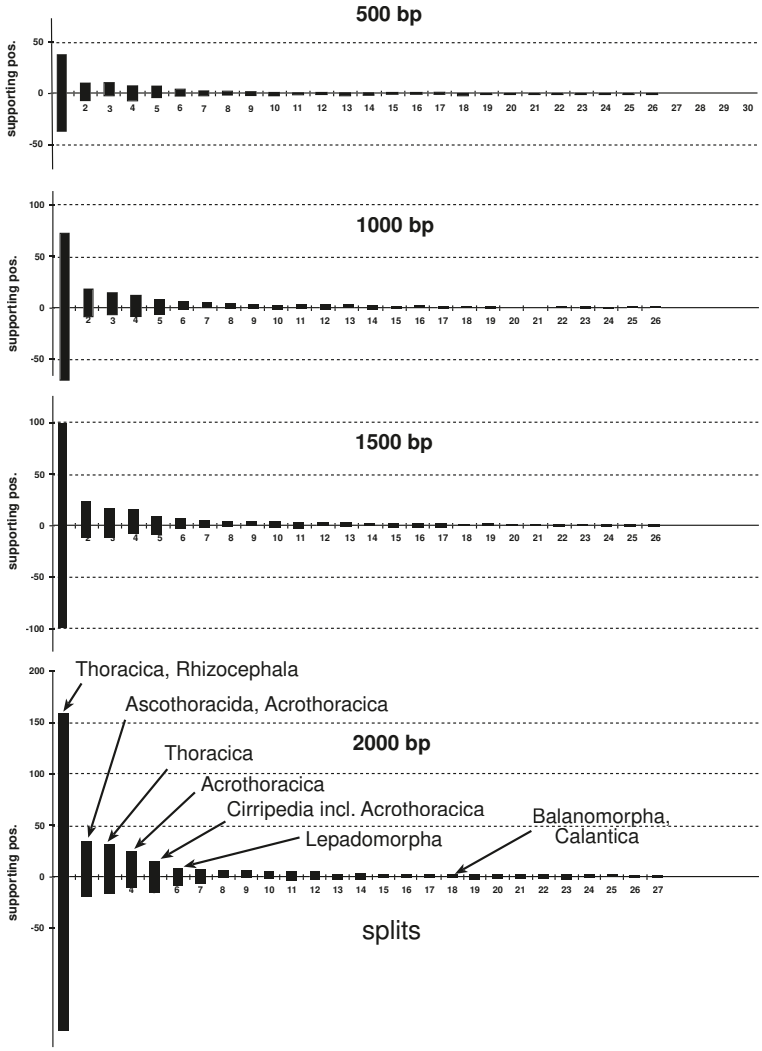


Fig. 154. Spectra of supporting positions showing the effect of alignment length (data from Spears et al. 1994). With increasing alignment, homology signal accumulates quickly, while noise is scattered over thousands of different splits. For those familiar with crustacean systematics: the split {Ascothoracida, Acrothoracica} is supported by homologies, however, these are plesiomorphic due to insufficient taxon sampling. Therefore, the group is paraphyletic.

– For each split S record the number n_A of sequence positions which support the split, when *A* serves as potential ingroup, but also the number n_B of sequence positions, when *B* is used as potential ingroup.
– Order the splits according to the number of split – supporting positions (Figs. 153, 170).

Note that the signal visualized with these spectra is a ***homology signal***. Whether the signal is based on apomorphies or plesiomorphies depends on taxon sampling.

This phenomenological method is still being developed and represents an attempt to transfer the methods of comparative morphology to the analysis of sequences. The decisive problem is the estimation of the probability for patterns: how much noise can be tolerated without inflating the pattern of supporting positions with symplesio-

morphies or analogies? If in a more conservative approach the number of split-supporting positions is limited to those positions that are less variable, information is lost and the support of many splits becomes too weak to allow the identification of potential monophyla.

Spectra are ideal tools for the explorative analysis of data. The visualization of the increase of homology signal with increasing alignment length (Fig. 154) shows why it is wise to work with concatenated sequences. Signal increases because noise is scattered over all possible splits of a data set, while sites with nucleotide patterns caused by conserved phylogenetic signal will support always the few splits that contain clades of the true tree. Often it can not be predicted which clades have the best support. The spectrum will show which monophyletic groups inferred by a phylogenetic analysis are the most reliable ones.

6.6 Combined analyses, data partitioning, total evidence

Combined analyses can be carried out for morphological data and molecular data, but also for a selection of sequences or for different genes.

When morphological as well as molecular datasets are available for a set of species, there are 3 alternatives to find a phylogenetic hypothesis:

a) *Separate analyses and plausibility test*: data are evaluated separately, the alternative dendrograms are tested in respect to their taxonomic congruence (see ch. 10). A tree is assembled with those monophyla that are mutually compatible and at the same time appear to be plausible.
b) *Separate analysis and democratic voting (consensus approach)*: data are evaluated separately and the results obtained from individual datasets are used to construct a consensus dendrogram, for example showing only those clades shared by all or by a majority of alternative topologies (ch. 3.3), or a supertree is constructed when species sets are only partly overlapping (ch. 3.3.1).
c) *Conditional combination*: only those data are combined that are not heterogeneous. Total evidence is strived for except when the data are shown to be incongruent.
d) *Combined analysis and addition of signal (total evidence)*: all available data are pooled to obtain a single data matrix to construct a dendrogram. Combinations of morphological and molecular data are analysed with the MP-method.

For (a) and (b) different phenomenological or modelling methods of tree inference can be used, adapting models individually to different types of data. For the analysis of combined morphological and molecular data ("total evidence") only the MP-method is suitable, because presently available modelling methods cannot consider different substitution models (for morphological or molecular data) at the same time for different parts of a data matrix. Furthermore, it is not known how to estimate model parameters for morphological data, and it might be true that morphological characters rarely evolve stochastically.

These alternatives are not equivalent. Since there is only *one* historically correct phylogeny, contradictions in dendrograms obtained with separate analyses of different data are caused (1) by differences of the information content of the characters used and by varying abilities of the methods to recover this information or (2) by different character histories (e.g., lateral gene transfer). Therefore, dendrograms based on different datasets do not necessarily have the same value. A prerequisite for the construction of a consensus diagram or a supertree (case (b)) is that different dendrograms have the same probability of being correct.

Differences in character evolution and in probability of homology

Shared complex morphological character states of real organisms have a higher probability of homology than single sequence positions. If both types of characters are entered individually and with equal weights in the same data matrix, an accidental or convergent nucleotide identity will neutralize the signal of an important morphological character if both characters support incom-

patible splits. The combination of data in one matrix (alternative (c)) is therefore dangerous: informative characters can become ineffective among the mass of noisy sequence data. As it is not known how to estimate and weigh the probability of homology of sequence data in comparison with morphological data, a greater objectivity can be obtained with a separate analysis (alternatives (a) and (b)). "Total evidence" does not necessarily produce the best tree, it only gives you the *most parsimonious compromise*.

Since there is reason to assume that datasets differ in their signal to noise ratio, one might argue that the separate analysis of the data will be safer, because a dataset with strong signal will with high probability support the correct tree, while the combination with weaker data will never be as good. It is like pouring a cheap wine into a Chateau Clerc Milon. This will be the case when morphological data based of complex character states are combined with short and noisy sequence data (the latter would be the cheap wine). In this case total evidence is a most parsimonious but less desirable compromise.

Total evidence

However, if the signal differences are not known or not obvious, and if there is no other evidence available that might help to identify the correct tree, addition of data to a large dataset is a promising approach *because homology signal accumulates much faster* than false signal caused by chance similarities (see Fig. 154: spectra). This is very obvious when alignments are expanded. And, if there is some bias in the evolutionary process, one might hope that this bias does not occur in all types of data. To get optimal results it is necessary that

- morphological characters are weighted according to their complexity (ch. 5.1),
- non-independence of characters (e.g., two characters arising from a single event) is compensated with lower weights,
- alignment areas of uncertain positional homology are eliminated,
- alignments are based on orthologous sequence regions,
- known differences in substitution processes are considered with weighting matrices.

The additivity of homology signal (Fig. 154) and the dilution of noise in a large number of splits is an *argument against partitioning*. A combination of two partitions of comparable quality that give incongruent results in separate analyses may gain phylogenetic support and resolution due to these effects (Gatesy et al. 1999). Differences in probability of homology in data regions can be adjusted by weighting when MP methods are used. Furthermore, partitions may contain signal for different time levels in the tree, so that a better resolution is obtained with combined data. Many partitions and criteria to identify partitions are possible, and one cannot avoid subjectivity (Chippindale & Wiens, 1994).

Anyway, in each case the plausibility of the results should be discussed.

Data partitioning

However, *separate analyses can be the better alternative* in some cases: using molecular data, adding different sequences to a large alignment can imply the combination of different substitution histories. To consider these differences in a model-dependent phylogenetic analysis it would be necessary to develop methods that allow the consideration of different models for parts of the tree and for regions of the alignment. Whenever such methods are not available, it is often recommended to partition the data into portions that have the same substitution history. Furthermore, due to paralogy of genes, incomplete lineage sorting or horizontal gene transfer different genes may have a different phylogeny and will therefore not fit to the same topology. The discovery of significant incongruence is relevant for the *conditional data combination*: one would combine only those data that are not too incongruent.

Example: separate analyses of one tRNA and five different rRNA genes of vertebrates led to different estimates of phylogeny. Four out of six genes recovered the phylogeny that is also supported by the fossil record, with a sistergroup relationship between crocodiles and birds, one gene gave varying results depending on the model used, and the 18SrRNA gene consistently united birds and mammals, a result that is clearly contradicting morphology and the fossil record (Huelsenbeck et al. 1996a). Total evidence analyses recover the clade birds + crocodiles, excluding the 18SrRNA gene the bootstrap support increases. Obviously, it was not pos-

sible to find a substitution model that describes correctly the evolutionary process for the 18SrRNA gene.

Some principles of data partitioning are explained in the following:

Often the available data are already partitioned due to independent data acquisition (anatomical studies, different sequencing projects, etc.). These data can be analysed separately considering different weight matrices or substitution models. However, if single long sequences contain areas that evolved under different selection pressures a single model may not explain simultaneously all parts of the dataset, while separate analyses using different partition-specific models could reduce incongruence. One might want to compare first, second and third codon positions, or stems and loops of a secondary structure, and it may be interesting to exclude a region of an alignment whose substitution history differs markedly from the rest. To find or to discern partitions, statistical tests can be used that indicate dataset incongruence. Incongruence can be measured as differences in the fit of data in separate or combined analyses (*character congruence*), or by comparison of the topologies of trees obtained from partitions and combined data (*topology congruence*).

Differences in branch support: a separate MP-analysis of partitions can show if there are incompatible clades with high bootstrap support. This would indicate partition heterogeneity. The same is true when by addition of some data to a large dataset the support for clades with high bootstrap values decreases.

Incongruence length difference (ILD): This test is based on the MP approach. The tree length obtained from single partitions is compared with that of the combined analysis (Farris et al. 1995). The combined dataset will have a higher proportion of homoplasies if the partitions support conflicting topologies. The significance of tree length differences can be tested with the *partition homogeneity test* (*PHT*): remove invariant characters. First add the tree lengths obtained in a separate analysis of each partition and subtract this sum from the tree length obtained for combined data. Then construct randomized partitions (with the same total length as the original data and the same number of partitions), with varying partition lengths and mixing characters of the original partitions. Count how often the length difference (the ILD value) obtained in the randomizations is the same or greater as in the original data (e.g., 20 of 1000 randomizations: $p < 0.02$). According to Cunningham (1997) the ILD test performs better than the following two ones.

Templeton's test compares the support for different topologies by a given set of characters. One can map the characters of a partition on two competing topologies to see how many characters fit to which topology. The number of character state changes in a most parsimonious state distribution on the topology is compared for each topology and significance of character support is estimated with a Wilcoxon signed-rank test (Templeton 1983). The test can be used to decide if two trees are significantly different from one another.

Rodrigo's test (a topology incongruence test) is also based on the MP method. Most parsimonious topologies obtained from a single partition are compared to calculate the symmetric distance (Penny & Hendy 1985), which is the number of species groups that appear in only one of the trees when two trees are compared. The mean symmetric distance for a partition is calculated bootstrapping each partition more than once, and finally this distribution is compared to the null distribution within each partition to determine if differences are significant (Rodrigo et al. 1983).

Maximum likelihood based tests:

If topologies differ because they are based on different datasets, we would like to know whether topology differences are due to random variations in substitution processes or due to non-random differences in sequence evolution. The *likelihood heterogeneity test* (Huelsenbeck & Bull 1996) can be used to compare the likelihood for all partitions (or for different datasets) and for a given topology with the likelihood when there is no topology constraint. The null hypothesis is that the same tree underlies all data partitions (likelihood L_1), while the alternative is that different trees and different rates describe the different data (likelihood L_2). The test statistics δ is calculated as $2(\ln L_1 - \ln L_2)$ The null distribution of δ is obtained from simulations (for details see Huelsenbeck & Bull 1996).

7. Process-based character analysis

Unlike phenomenological analyses, model-dependent methods do not estimate *a priori* how probable it is that patterns of identical details can be identified correctly as being homologies ("probability of cognition"), but instead estimate the probability that two patterns (whether they are similar or not) evolved from a common ancestor pattern, i.e. the "probability of event". The "event" is the process of evolutionary modification of a character or the evolutionary assemblage of a novelty. Ideally, the expected frequency of specific evolutionary events has to be estimated to be able to calculate the probability that identities are not products of chance (analogies) but the result of common ancestry.

When weighting according to the probability of homology, character weights are reciprocal to the expected frequency of character evolution (Fig. 155), because frequent events produce identities by chance more often than rare events. In order to estimate the expected frequency it has to be known (or to be assumed correctly with some certainty) which process produced the character states in terminal species. The expected result of the process has to be described quantitatively.

For the correction of genetic distances, the probability for the occurrence of specific substitutions per unit of time is considered. The statistically more frequent events contribute more to the occurrence of multiple substitutions. The latter reduce the visible distance (number of differences) between two sequences compared to the real (evolutionary) distance.

All modelling methods rely on the axiomatic assumption that character evolution is predictable, implying that it is a stochastic process.

Morphological characters

In practice, the probability of homology of morphological characters is mostly determined phenomenologically (see ch. 5), because parameters of evolutionary processes that cause the modification of morphological characters (e.g., the number of mutations per unit of time, effects of selection processes) are not known. At present the processes cannot be simulated realistically for those cases that are relevant for systematics, or they are not recordable quantitatively. When assumptions on processes are part of the argumentation, then usually without definition and estimation of process parameters, so that in the end the statement of homology is substantiated phenomenologically.

Example: in the genus *Cylisticus* (terrestrial woodlice), species normally live on the surface of the soil (epigeic, group *A*), some smaller ones live

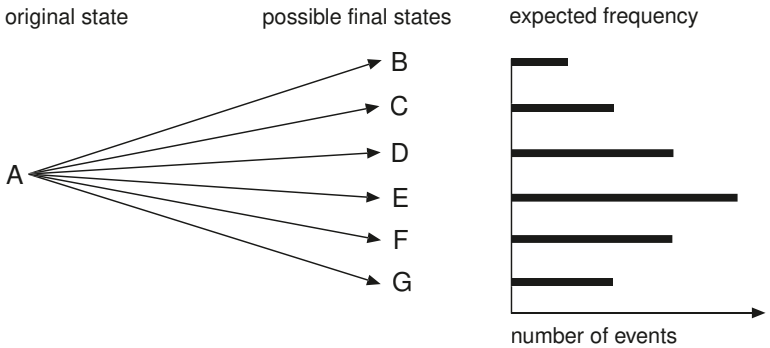


Fig. 155. Model-dependent weighting of characters based on the estimation of evolutionary processes: the most frequent event has to get the lowest weight, because here the probability is greatest that similarities occur by chance alone. Rare events produce better homology signals, they are conserved over longer periods of time.

subterraneously (endogeic, group *B*). It is obvious that one can assume that the endogeic group descended from an epigeic ancestor. The following fact is observed: (1) species of group *B* (endogeic) have complex pleopod lungs. (2) Species of group *A* do not have such complex lungs. A process assumption is coupled with this observation: it may be expected that as an adaptation to the subterraneous way of living, body size was reduced and therefore also the lungs should be simplified. It is at first sight surprising that lung function was enhanced instead. It is concluded that species of group B did not evolve from an ancestor which belongs to species group A or that the ancestor did not have the same structure of pleopods as in group A. Rather there must have been a common ancestor with already well developed complicated lungs. The expected reduction of lungs did not or not yet occur in group B (from Schmidt 1999).

In this argumentation a statement on the origin of group *B* is linked with an assumption about a process (the probability of the evolution of lungs in endogeic habitats). A statement of homology, however, is not combined with this hypothesis about processes. The homology of the lungs of group *B* has do be inferred from structural identities. Furthermore, it is apparent that a model with a quantitative statement about the probability of lung evolution cannot be proposed. The argument also implies that evolution of lungs should be a slow process and is not expected to happen within a group of closely related species. Assuming that lungs evolved faster and that they are needed underground where oxygen concentration is low, then a different hypothesis can be supported, namely evolution of group *B* from an ancestor with the morphology seen in group *A*.

Process assumptions are usually not very useful for analyses of the homology of morphological characters due to lack of information about the real evolutionary process and about probabilities of character evolution, even though in theory character weighting could be improved. Character co-variation due to function may indicate that functional correlation increases the probability of analogous change.

In general, characters can show a *phylogenetic covariance* because they change on the same branches of a tree, while *functional covariance* is explained by a functional "cooperation" of characters in the same individual. Functional covariance may occur on the same branch of a tree (a functional complex is a synapomorphy), but it can also evolve in parallel on different branches. Functional correlation can occur at different levels: correlation of genes during development, pleiotropy, physiological or behavioural correlation, and biomechanical correlation. If a new function is an adaptation to environmental parameters, convergent adaptation can also lead to a parallel change of several characters (Emerson 1998). How to consider these phenomena to weigh morphological is discussed in chapter 5.1.

Assuming that probabilities of character state changes can be described with a model, it is possible to use maximum likelihood methods to construct phylogenetic trees. An example for a set of assumptions (Lewis 2001): the probability of character state changes is symmetrical (equal probability for gains and losses and for any substitution) and constant along a branch between two nodes, wherefore the probability increases with branch length (which does not exclude a punctuated equilibrium on a single branch because the rate is an average value); probabilities are the same for each character state (equally weighted character states).

DNA-sequences

Each mutation of a sequence of an organism is an evolutionary novelty. However, after several generations not every mutation is present in a population (Fig. 5). Assuming that two homologous sequences are evolving *independently* in two separate populations with the same regular, low mutation rate, and that this rate is the same for all nucleotides, and that all nucleotides have the same frequency, and if selection has in both populations the same effect, the following results are expected comparing sequences from both populations:

- some mutations occur in single sequences (autapomorphies),
- some substitutions produced analogies (chance similarities) shared by two homologous sequences of different populations,
- some mutations produced character states identical to a previous state ("back mutations").

Taking the conditions listed above to estimate statistically the frequency of substitutions and comparing two homologous and parallel evolving sequences, autapomorphies occur after each mutation (probability 1), chance similarities (when different nucleotides were present before) appear after each third mutation (probability 0.333), and back mutations after each fourth mutation (3 of 12 mutations, probability 0.25) (Fig. 156). Note that some true autapomorphies have the appearance of analogies, others are back mutations. These autapomorphies are therefore "invisible".

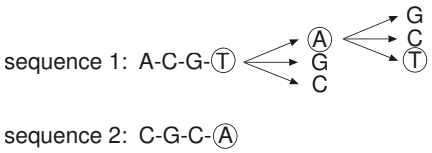


Fig. 156. Possible mutations for position 4 in sequence 1 and the appearance of analogies (in relation to position 4 of sequence 2) and back mutations.

When a mutation spreads in a population due to genetic drift and/or selection, it can become a substitution. Considering short periods of evolution and low substitution rates, multiple substitutions (repeated substitutions at the same sequence position) are rare and therefore character states shared by related organisms are expected to be with greater probability apomorphic homologies than the result of substitutions which cause signal-like "noise" (from the point of view of phylogeneticists these are analogies and back mutations). Autapomorphic substitutions in populations become synapomorphies after speciation events, wherefore under favourable conditions, that is when a number of unique substitutions is inherited from a common ancestor and the probability for accumulation of analogies and for multiple substitutions is low, shared identities represent with greater probability homologies than analogies. Of course, synapomorphies can only be identified when different character states are present in the outgroup.

For this reason the individual substitution is only informative when no or only a few multiple substitutions are expected to occur. Over longer periods of time the individual substitution is irrelevant for the analysis, and the probability of homology of a single shared nucleotide is low. To identify homology signals, substitutions of many sequence positions have to be considered and the probability that multiple substitutions and analogies evolved has to be estimated. These probabilities are obtained with the help of models of sequence evolution (ch. 8.1), which are used in distance- or maximum likelihood-methods (see chapters 8.1, 8.3, 8.4). Alternatively, spectra (ch. 6.5.2) can be constructed to visualize the frequency of chance similarities.

The individual character is only evaluated as part of patterns contained in the complete dataset, for example by calculating genetic distances with the help of a selected model of character evolution, or in order to estimate the probability that a given dataset could be the result of a chosen model of the evolutionary process along a given topology. These methods are explained in chapter 8 and in the more detailed paragraphs of the appendix (ch. 14). In distance methods, characters shared by two sequences (or differences between them) are counted and the numbers are transformed with models for the substitution processes to obtain the real (evolutionary) distances. With maximum likelihood methods individual characters are considered and the probability that the character transformation from some ancestral character state to its descendants fits to a given topology is estimated using a selected model of sequence evolution.

8. Reconstruction of phylogeny: model-dependent methods

These methods were developed to reconstruct phylogenetic trees using well defined statistical approaches. This endeavour requires an **axiomatic assumption:** the evolution of characters is a **stochastic process** that can be described or simulated with models. This means that on average the same set of substitution probabilities is constant for defined types of substitutions in a defined region of a gene over long periods of time and in different types of organisms. It is presupposed that each sequence position evolves in a way that allows the description and prediction of evolutionary changes of a sequence or gene with a mathematical model, and without testing this assumption, the main focus lies on the selection of optimal model parameters.

Each model can only simulate processes and is not *a priori* a reliable copy of nature itself. Experience teaches that models only contain those variables which are conspicuous and regular and therefore can be considered. For this reason simulated evolutionary processes are generally much simpler than real ones. Furthermore, historical processes can only be modelled when they left traces or when the scientist believes he or she can find some clues for the real course of the processes. This can be problematic especially with evolutionary processes. There is the danger that errors and unfounded assumptions (*ad-hoc* hypotheses) influence the model decisively. On the other hand, it is possible with model-dependent methods to homologize patterns that show only few similarities, a case where phenomenological methods do not recover sufficient information to justify a decision.

Axioms necessary for the use of model-dependent methods:

- the evolution of the characters used for the analysis is a stochastic process
- the model does not deviate much from the real historical processes
- the available data are representative for the historical events

8.1 Substitution models

Complex probabilistic models have been developed especially for character transformations in DNA sequences. In distance methods they serve to estimate the number of non-observable multiple substitutions, and in maximum likelihood methods to estimate the probability for the occurrence of specific substitutions and ground patterns (node characters). The formal description of simple models is presented in the appendix (ch. 14.1), their application in distance methods is discussed in chapters 8.2 and 14.3, for maximum likelihood methods see 8.3, 8.4 and 14.6.

A basic idea for the description of a model of sequence evolution is the assumption that for specific substitutions a certain rate exists that describes the average number of substitutions per unit of time. The model assumptions can then be described in a rate matrix (Figs. 157, 158). Rates are primarily relative, i.e. without a specification of the unit of time. For the Kimura-2-parameter-model (K2P), for example, two rates are distinguished (α for transitions, β for transversions, compare Fig. 157). When nucleotide A is found at a position of a sequence, than it may have replaced a C or a T in the ancestral sequence with the rate β or a G with the rate α. The rate thus takes the role of an assumption about probabilities: when a specific rate is higher than others, the respective substitution will occur with higher probability per unit of time. The rates are standardized in such a way that on average one substitution occurs per unit of time; therefore, in the K2P-model we get $\alpha+2\beta=1$ (see Fig. 42). In the **rate matrix** the expression $(-\alpha -\beta)$ is used as "rate" (probability) for the unchanged nucleotide.

Based on these considerations a matrix can be compiled for each model which describes how probable it is that a nucleotide *j* evolves from a

nucleotide *i* in the time interval *dt*. In some models that have been popular until now further parameters are incorporated, especially the frequency of nucleotides found in alignments and the frequency distribution of different rates (see below and ch. 14.1). However, it is not possible to consider real variations of selection constraints which can cause changes of substitution rates along a single branch and along different parts of a phylogenetic tree. It is only possible to count, for example, the Ts:Tv-ratio (K2P-model) in pairwise comparisons of the available sequences in order to use this value in models. This procedure requires the assumption that selection always has the same effect during the course of time. This assumption, however, is usually not discussed and it is in most cases probably not testable.

Models of sequence evolution imply assumptions which have to correspond to the historical reality whenever the model is required to reconstruct phylogeny. One assumption which is often expected to be valid is for example that the substitution rate is independent of the region in the molecule and independent of preceding events. Whereas the first assumption does often prove to be false, the second one is probably correct in most cases: in the series of substitutions $C \rightarrow T \rightarrow A$ the probability that T is replaced by A is independent of the fact that historically older populations possessed in place of the T a C. More complex model-specific assumptions on substitution rates can be described with a substitution matrix (Fig. 158).

In this matrix (Fig. 158), the diagonal contains an expression for the probability that the nucleotide remains unchanged. The value has been selected in such a way that the sum of a row of the matrix is 0. The expressions $\alpha_1\pi_c$ means that the probability of the substitution $A \rightarrow C$ depends on the specific rate α_1 and on the frequency π_c of the nucleotide C in the real sequences.

	A	C	G	T
A	-3λ	λ	λ	λ
C	λ	-3λ	λ	λ
G	λ	λ	-3λ	λ
T	λ	λ	λ	-3λ

	A	C	G	T
A	(-α-2β)	β	α	β
C	β	(-α-2β)	β	α
G	α	β	(-α-2β)	β
T	β	α	β	(-α-2β)

Fig. 157. Rate matrix for the Jukes-Cantor model (top) and for the Kimura-2-parameter-model (bottom). The models are reversible, i.e. independent of the direction of a substitution ($A \Rightarrow C = C \Rightarrow A$).

The model of Fig. 158 has a defined time axis and therefore is not reversible, because the rate for $A \rightarrow C$ does not have to be the same as the rate for $C \rightarrow A$. Reversible models have the same rate independently of the direction of the substitution for any substitution between two nucleotides (rate for $A \rightarrow C$ = rate for $C \rightarrow A$) and can be arranged in a diagonally symmetrical matrix (Fig. 157).

Generally, models make assumptions on:

- the probability that a specific nucleotide is replaced by another nucleotide;
- the substitution rate per position: positions can evolve faster or slower independently of the type of substitution, the speed can be uniform for the whole sequence or vary substantially within the sequence, a portion of the sequence can be invariable;
- the homogeneity of sequence evolution in the course of phylogeny: the model of evolution holds for all stem lineages of a phylogenetic tree. Today there are no modelling methods which do not require these assumptions. However, it is very probable that molecular evolutionary processes vary in time and with different species (non-stationarity of processes) and that the number of variable positions is different in different taxa.

	new character state			
	A	C	G	T
A	$-(\lambda_1\pi_C+\lambda_2\pi_G+\lambda_3\pi_T)$	$\lambda_1\pi_C$	$\lambda_2\pi_G$	$\lambda_3\pi_T$
C	$\lambda_4\pi_A$	$-(\lambda_4\pi_A+\lambda_5\pi_G+\lambda_6\pi_T)$	$\lambda_5\pi_G$	$\lambda_6\pi_T$
G	$\lambda_7\pi_A$	$\lambda_8\pi_C$	$-(\lambda_7\pi_A+\lambda_8\pi_C+\lambda_9\pi_T)$	$\lambda_9\pi_T$
T	$\lambda_{10}\pi_A$	$\lambda_{11}\pi_C$	$\lambda_{12}\pi_G$	$-(\lambda_{10}\pi_A+\lambda_{11}\pi_C+\lambda_{12}\pi_G)$

Fig. 158. Substitution matrix with model-specific substitution rates. In this example 12 different substitution rates λ are distinguished, and the base frequency π is considered.

substitution probabilities do not change with time	+	+	+	+	+	+	+	+
all substitutions are independent of each other	+	+	+	+	+	+	+	+
sequence evolution is homogeneous	+	+	+	+	+	+	+	+
the direction of substitutions is not relevant (model is reversible in time)	+	+	+	+	+	+	+	+
base frequencies are constant in time	+	+	+	+	+	+	+	+
sequence evolution is stochastic	+	+	+	+	+	+	+	+
base frequencies are equal (1:1:1:1)					+	+	+	+
6 classes of substitution rates can be distinguished	+				+			
3 classes of substitution rates can be distinguished (transitions, 2 classes of transversions)		+				+		
2 classes of substitution rates can be distinguished (transitions and transversions)			+				+	
there is only one substitution rate for all events				+				+
model:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

Fig. 159. Table with assumptions of substitution models; models are named with commonly accepted abbreviations (after Swofford et al. 1996, with additions). For a more detailed table of models see Fig. 160.
1: GTR ("general time reversible model": Lanave et al. 1984, Tavaré 1986, Rodriguez et al. 1990)
2: TrN (model by Tamura & Nei 1993)
3: F84, HKY85 (models by Felsenstein 1984, 1993; Hasegawa-Kishino-Yano-Model of 1985)
4: F81 (model by Felsenstein 1981)
5: SYM (model by Zharkikh 1994)
6: K3ST (3-substitutions-model by Kimura 1981)
7: K2P (2-parameter-model by Kimura 1980)
8: JC (model by Jukes & Cantor 1969)

Attention: The improvement of models by addition of parameters does not imply at the same time an improvement of the reconstruction of phylogeny! Many of the popular models are very simplified. The more complex the models are and the more parameters are considered, the smaller is the number of strict assumptions that are required, but also the higher is the dependence of the correctness of the estimated model parameters (Waddell & Steel 1997, Philippe & Laurent 1998).

The selection of the model lies with the systematist, there exist however methods that help to identify models that fit to a data set (likelihood ratio test: ch. 8.1; Bayesian analysis: ch. 8.4). Often it is not exactly known which axiomatic assumptions are implied with a model. The table (Fig. 159) contains some important model-specific assumptions required for the analysis of DNA-sequences.

Several improvements of these models exist, for example, to consider the heterogeneity of substitution rates per site (ch. 14.1.5). The dependence of positional variability on secondary structure must be explored empirically. Models considering the chronological and lineage-specific variability are computationally complex and sometimes intractable. However, despite the imperfection of models, even the use of the simple Jukes-Cantor model is preferable to the omission of any corrections.

There exist invariable positions in each phylogenetically interesting sequence alignment. Their number can differ between species groups. Positions can be conserved in some species, but be subject to a reduced selection pressure and show more frequent mutations in other taxa. For a specific gene, the substitution rate in sequence regions can be very different (see Fig. 46) and can vary in time. When this is not taken into account, mistakes can occur in phylogeny inference. When using model-dependent methods, it is not possible to omit explicit assumptions about the molecular evolutionary process.

Some of the axiomatic assumptions are independent of the type of model. It must be assumed that

- the alignment contains the correct positional homology;
- the species sample selected for a taxon is representative for this taxon;
- the analysed sequence region does not contain sampling errors (is representative for the average type and frequency of substitutions and free of sequencing errors);
- in principle, the evolution of the sequence is a stochastic, predictable process;
- the selected model is applicable for all sections of the phylogenetic tree.

Some of these assumptions cannot be considered by the selection of a suitable model, but only by a control of the alignment, the selection of suitable species, and by use of long and informative sequences (see symplesiomorphy trap: ch. 6.3.3, influence of sequence length: ch. 9.1). The assumption of the existence of stochastic substitution processes always remains a paradigm and risk.

Currently available substitution models generally allow no statements on the probability that insertions or deletions occur. Character columns (positions) of a data matrix in which gaps occur are therefore ignored. This implies the assumption that these changes in sequence length are neutral for selection processes in neighbouring sequence areas.

The concepts implied in a model can be expressed with formulas which allow the deduction of probability statements of the following kind: assuming that a specific substitution rate α exists for a time span t, the character state A can be replaced with higher probability by state B than by state C. (Reminder: in DNA sequences the character (frame homology) in the sense of these models is the sequence position, the state (detail homology) is the specific nucleotide.) In the simple **Jukes-Cantor** model (see ch. 14.1.1), for example, for the probability P_{ij} that the nucleotide i is substituted by the state j we get:

for i=j (no change, conserved sequence position)

$$P_{ij}(t) = \frac{1}{4} + \frac{3}{4}e^{-\lambda t}$$

for i≠j (substitution occurs)

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-\lambda t}$$

In these formulations the **substitution rate** λ and the time interval always appear as product $\lambda \cdot t$, which corresponds to the **branch length** when branch length is defined as the **number of substitutions** which occur along a branch. It is intrinsically not possible to separate rate and time without additional information. Fortunately, it is not necessary to know either the absolute value of the substitution rate (e.g., number of substitutions per one million years), nor the real divergence time between terminal species. It is completely sufficient to determine the relative branch length in order to draw a dendrogram. This can be achieved with maximum likelihood and distance methods.

In order to be able to calculate a probability for a topology (compare ch. 8.3: maximum likelihood method, ch. 8.4: Bayesian phylogeny inference), the parameters used for the models have to be estimated, for example, the base frequency or the putative branch length $\lambda \cdot t$.

The **base frequency**, for example, can be estimated by pairwise comparisons of terminal sequences, or for a total dataset by calculating the average of the base frequencies visible in single sequences of an alignment. This procedure requires that the ancestral sequences in nodes of the tree also contained the same average base frequency. A further variable which can be included in *all* methods, is the **positional variation of the substitution rate**. To use this parameter for the complete alignment it is necessary that the sequence positions of a macromolecule are classified according to their variability. An estimation of this variability can be derived from the observation that some positions have the same character state in all sequences, whereas others show differences more or less often. In the simplest classification, two classes of positions can be distinguished: constant and variable ones. The other extreme is a continuum of variants of the substitution rate, visualized in a frequency distribution curve (number of positions plotted against rates) which shows how often sequence positions show individual rates (see gamma distribution and other models in ch. 14.1.5).

starting model	equal base frequencies	go to
JC	yes	Table A
	no	Table B

Table A: models with equal base frequencies

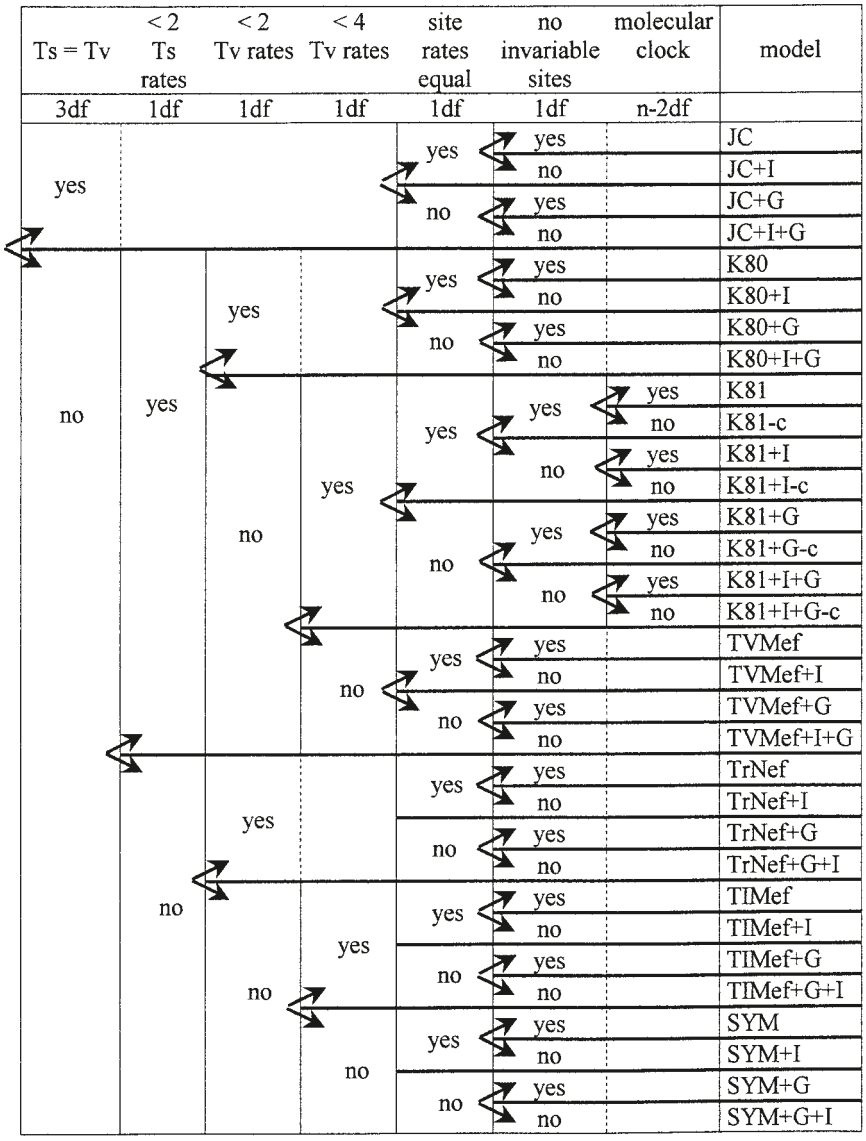


Fig. 160a. Hierarchy of models used for a likelihood ratio test (part A).

The **substitution rate** or the **branch length**, however, remain unknown and can only be estimated indirectly. There are methods ("**model-fit**"), which enable an optimization of parameter values during maximum likelihood calculations (summary in Swofford et al. 1996, implemented in the program PUZZLE, see Strimmer and von Haeseler 1996, Strimmer 1997). The optimization again is

Table B: models allowing unequal base frequencies

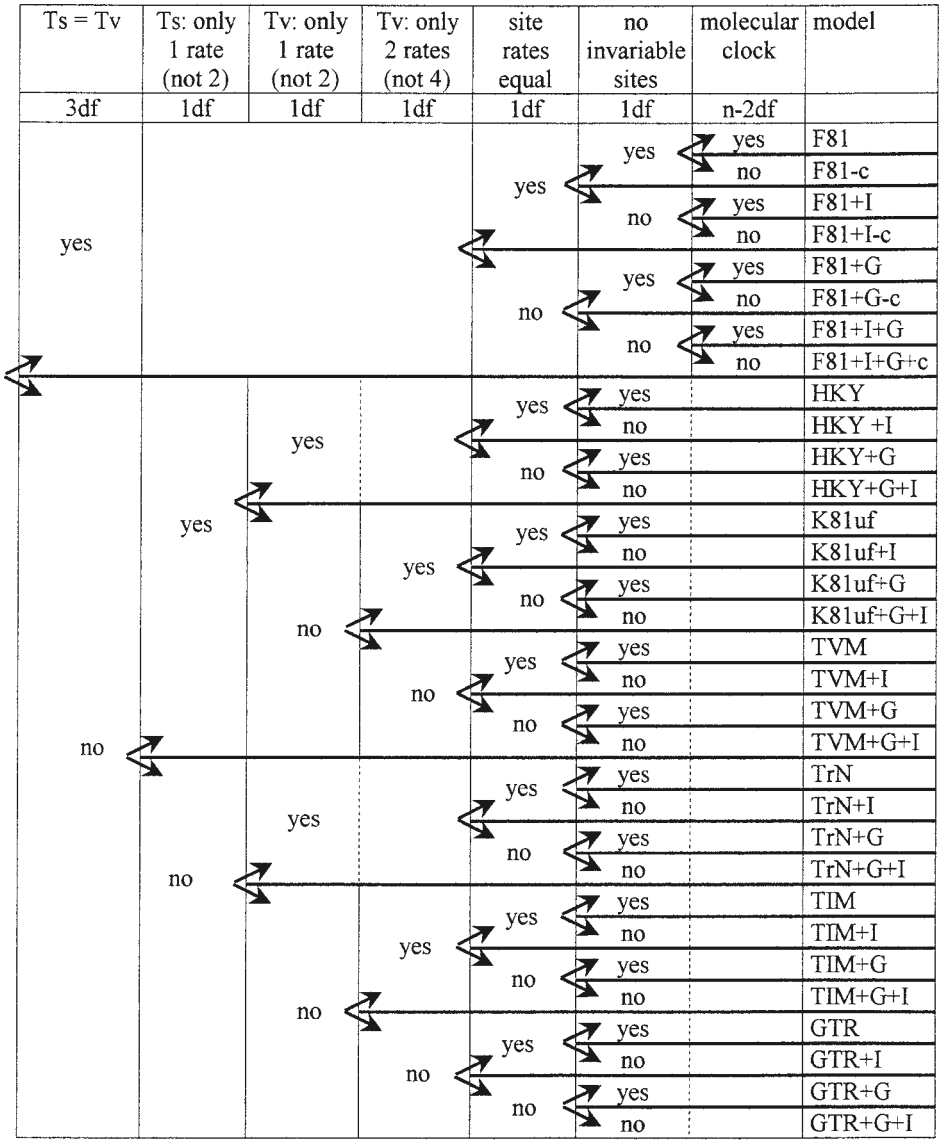


Fig. 160b. Hierarchy of models used for a likelihood ratio test (part B). Empty fields indicate that a model parameter is not applicable. The degrees of freedom (*df*) are the difference between the answers "yes" or "no" and depend on the number of additional model parameters required by a subordinated model. Model names: c (with enforced molecular clock), F81 (Felsenstein 1981), G (with gamma distribution), GTR (general time reversible model), HKY (Hasegawa, Kishino, Yamo 1985), I (with invariable sites), JC (Jukes & Cantor 1969), K80 (Kimura 1980), K81 (Kimura 1981), K81uf (K81 and unequal base frequencies), SYM (model of Zharkikh 1994), TIM (transitional model), TIMef (TIM and equal base frequencies), TrN (Tamura & Nei 1993), TrNef (TrN and equal base frequencies), TVM (transversional model), TVMef (TVM and equal base frequencies) (compiled by Posada & Crandal 1998).

model-dependent and there exists no method to test the correctness of the estimated parameters *a priori*. A direct "measurement" of process parameters is not possible. The plausibility of the "best" topology calculated to represent a phylogeny can only be tested with data of other origins (ch. 10).

The user is faced with the problem to **select the best model**. Theoretically, it is possible to choose models of sequence evolution that are so specific and complex that for a given dataset the probability of a specific topology using such a model is 1. For the same set of data another topology would get the same probability with a different complex model. Therefore, a wrong hypothesis of phylogeny can be supported with a model that is "too complex". The principle of the most parsimonious explanation has also to be applied in this case: the more *ad hoc* assumptions are implied, the lower is the probability that the hypothesis corresponds to the reality that exists outside our minds. With the "**Likelihood Ratio Test"** (Goldman 1993a,b) it can be tested whether the selection of additional model parameters causes a significant improvement of the probability for a topology. When this improvement is not significant, a simpler model should be retained.

For the "Likelihood Ratio Test" a topology and an alignment are given. L_0 and L_1 are the probabilities estimated with an ML-calculation for the given topology under model 0 and model 1, respectively. The test statistics β for the difference between the two probabilities is calculated with: $\beta = -2 \log (\max L_0 / \max L_1)$, whereby a χ^2-distribution is assumed. The number of degrees of freedom corresponds to the number of additional parameters of the more complex model. (A computer program for this test is available at http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm.)

A further possibility to fit models is the variation of model parameters and the selection of models and topologies according to their posterior probabilities (see Bayesian analyses, ch. 8.4).

When the optimal fit of a model is found with the maximum likelihood method based on a previously selected topology, and subsequently a ML-topology is calculated with the same model, the argumentation may become **circular**: model and topology are adjusted to the data, an independent test of data quality and of the real historical substitution processes is lacking. The only statement that is justified is that one has found a model that explains a topology with the given data.

Note that model parameters can be classified into **structural parameters** that appear in the likelihood function of all characters, and **incidental parameters** that are considered for only part of the characters. To get statistically consistent models the latter should be avoided (Lewis 2001).

Attention: even when a model has been selected carefully with likelihood ratio tests, it might still be that it fails to capture the relevant information about molecular evolution, or it might be that the data do not contain the information that is necessary to describe the evolutionary process.

8.2 Distance methods

A distance between two species should represent the distance in time since divergence from the last common ancestor, i.e., the time elapsed since the speciation event that separated the two lineages to which the two species belong. With distance methods, an optimal topology fitting to all data under the criteria of the selected method is constructed without having to test the support for individual putative monophyla with the given discrete characters. Such a test can follow afterwards, for example, using the bootstrapping method (ch. 6.1.9.2).

In principle, dendrograms can be obtained with different approaches:

- searching clusters of most similar sequences based on pairwise distances between sequences (*clustering methods*),
- seeking the tree whose sum of branch lengths is minimized (*minimum evolution methods*, see ch. 8.2.7).

In the following we will focus on *clustering methods*. Cluster analyses always yield some tree graph. There exist, however, less liberal methods which

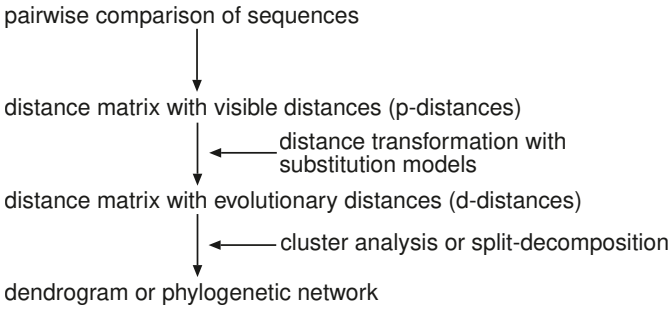


Fig. 161. Flowchart for a distance analysis.

produce a tree only when the data fit *unequivocally* to such a topology, otherwise networks are obtained (see split-decomposition, chapters 6.4, 14.4).

Distances between two species or between two ground patterns can be calculated for any character set when an objective numerical measure exists for the similarity or dissimilarity. Such a measure can, for example, be obtained for proteins with immunological methods (ch. 5.2.2.5), or for extracted DNA with DNA-DNA hybridization (ch. 5.2.2.8). The comparison of distances for discrete morphological characters, as it has been practiced in the early times of numerical taxonomy, can only yield a plausible result by chance, because it is generally not known which relationship exists between morphological differences and divergence time. The irregularity of the evolution of morphological characters (see ch. 2.7.1) can hardly ever be modelled over a larger time scale (a problem shared with weather forecasting), wherefore the approach of numerical taxonomy had to be given up.

Most frequently, however, these methods are used for the analysis of sequence data. The basis of distance methods is the comparison of pairs of aligned sequences. Each position of the two sequences is examined and the differences in character states are counted and added to a distance value (see below, ch. 8.2.2). Topologies representing phylogenetic relationships can be calculated with these data. Additionally, the following statistics can be obtained on individual sequences or on the alignment:

– frequency of the nucleotides of a sequence (relation A:T:C:G)
– frequency of codons in protein-coding sequences
– number of variable and conserved positions of an alignment
– frequency of paired nucleotides in the alignment of two sequences, whereby simultaneously transition differences $\boldsymbol{T}_s$ (A⇔G, T⇔C) and transversion differences $\boldsymbol{T}_v$ (A⇔T, A⇔C, T⇔G, C⇔G) can be counted to determine the relation $\boldsymbol{T}_s$:$\boldsymbol{T}_v$. Note: this is the visible difference, but not necessarily the real number of historical substitutions that separate the sequences from the last common ancestor.

The estimation of genetic distances is useful for

– the estimation of evolutionary rates,
– the estimation of divergence times.

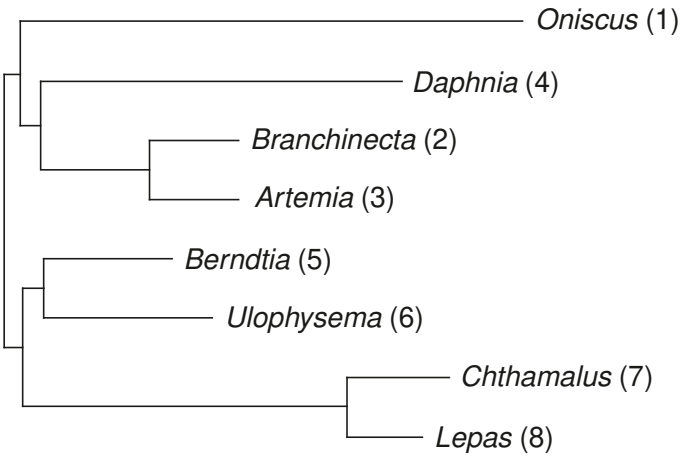
Distance methods allow a reliable reconstruction of phylogeny **only if** the distance values between pairs of species or sequences are a **measure of the divergence time**.

8.2.1 The principle of distance analyses

Distance trees are constructed without time-consuming tree search (in contrast to maximum parsimony or maximum likelihood methods), because a **fast algorithmic approach** is used.

Distance analyses are performed according to the following principle (Fig. 161, see Fitch & Margoliash 1967):

– Choose two aligned sequences and count the positions with unequal nucleotides in both sequences. The visible distance is the sum of the differences expressed as portion of the alignment length.



OTUS	1	2	3	4	5	6	7	8
1		0.2338	0.2442	0.2837	0.2383	0.2480	0.3039	0.2937
2			0.0585	0.1883	0.1283	0.1370	0.2249	0.2138
3				0.1797	0.1319	0.1351	0.2235	0.2083
4					0.1806	0.1922	0.2913	0.2788
5						0.0962	0.1806	0.1737
6							0.2018	0.2026
7								0.0580
8								

Fig. 162. Distance dendrogram for taxa of Crustacea, calculated from a 18S rDNA-alignment (2355 bp). The corresponding distance matrix shows that the smallest distance values occur for sequence pairs which are also neighbours in the dendrogram (2+3, 5+6, 7+8). (Neighbour-joining algorithm and Tamura-Nei-distance as implemented in the program MEGA (Kumar et al. 1993); positions with alignment gaps deleted in pairwise comparisons. OTUS = terminal taxa).

- Find the visible distances for all sequence pairs and construct a data matrix (Fig. 162). For n sequences there are $\boldsymbol{n(n-1)/2}$ distance values.
- Choose substitution parameters (a model for the evolutionary substitution process) to convert the visible distances into evolutionary distances.
- Choose a clustering method to calculate a dendrogram based on the distance matrix.

The distance transformation with substitution models is necessary to estimate the number of invisible multiple substitutions (ch. 2.7.2.4, Fig. 165). The larger the divergence time between two species, the more frequent are multiple substitutions and the larger is also the difference between the visible and the real (evolutionary) distance. In practice, however, the complex transformations of distances often prove to be less effective than originally hoped for. It can be shown empirically that most dendrograms maintain their topology independently of the selected evolutionary model used for the distance correction. When the topology is not plausible or obviously wrong, the sources of errors must be sought elsewhere. The cause may be, for example, the lack of phylogenetic signal, or incomplete species sampling, or a bias in base frequencies not considered in the model and causing a large number of chance similarities.

Distance methods are phenetic methods. In principle they do not distinguish between classes of characters (plesiomorphies, autapomorphies, analogies, synapomorphies) on all hierarchical levels (see ch. 4.2, Fig. 164, 165). Even autapomorphies, which are **trivial characters**, influence the estimated distances (Fig. 166). With very large amounts of data it is to be expected that randomly distributed chance similarities of sequences in an alignment do not produce clear relationships based on sequence similarity. They should remain "background noise" without effects, whereas the non-random homology signals (Fig. 154) decide on the topology of the dendrogram. With

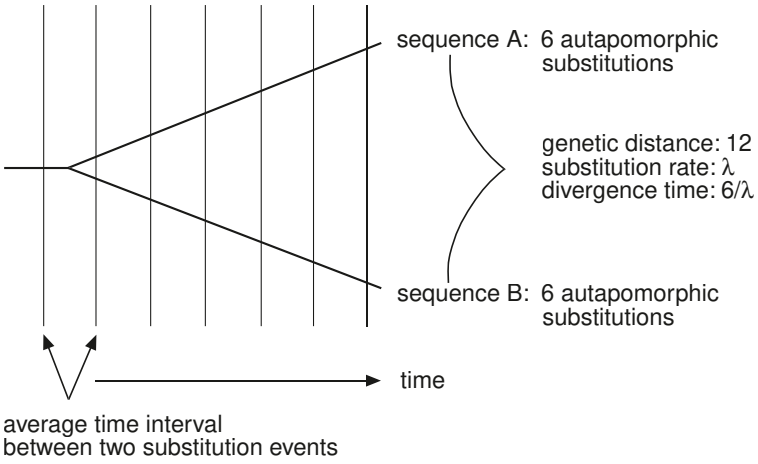


Fig. 163. Relation between divergence time and genetic distance when the substitution rate is constant (= constant "molecular clock") and multiple substitutions do not occur. The distance of the sequences (the path between A and B) corresponds to $2\lambda t$, the branch length between the basal and the terminal nodes is in each case λt.

smaller amounts of data chance similarities may accumulate in some groups and give false evidence for monophyly.

In principle, sequences cannot be evaluated for a phylogenetic analysis when all or the majority of the variable positions of a sequence are substituted in pairwise comparisons more than once on single branches (saturated sequences; see Figs. 39, 43). This holds also for distance methods even when the effect of multiple substitutions is considered with models. Note that even a few multiple hits destroy homology signal when a sequence has only a few variable positions and the overall distance between the sequence is small due to the presence of many invariable positions.

Contrary to parsimony or likelihood methods, distance methods have the advantage that calculations are very fast. Distance methods require the assumptions that the real historical course of evolutionary processes (mutations, fixation of mutations in the population, effects of genetic drift and of population bottle necks, occurrence of multiple substitutions, fluctuations of substitution rates) that produced the observed distances can be described statistically with models of sequence evolution. As the variance of these processes increases with larger time scales and with increasing genetic divergence between species, a distance analysis based on single genes is not a trustworthy method for phylogeny inference when large divergence times have to be considered with only a few gene sequences.

8.2.2 Visible distances

An essential basis for the estimation of divergence times and distance topologies is the calculation of the number of real substitutions per unit of time for a specific sequence region. This unit can be a relative measure, its absolute value is not relevant as long as we are only searching a topology. It should allow a statement on the relative divergence times between different organisms or – if branch lengths represent time – branch lengths should be proportional to the real age of lineages.

Whenever real divergence times between two organisms can be determined as in Fig. 163, distances are said to be **ultrametric**. These distances are directly proportional to the historical divergence time (ch. 14.3.4). Perfectly ultrametric distances have the advantage to fit unequivocally to only one rooted dichotomous dendrogram. When lineages are found that evolved with a constant molecular clock as in Fig. 163 and assuming that chance similarities are either absent or corrected with a model of sequence evolution, it is possible to obtain a phylogenetic tree from distance data using simple clustering methods (UPGMA: see ch. 14.3.7). However, when substitution rates are not constant, the distances may be at least **addi-**

tive (see appendix 14.3.3), in which case they fit directly to exactly one unrooted topology. However, biological sequence data usually are not perfectly additive, for example, because chance similarities produce contradictions. The neighbour-joining algorithm (ch. 14.3.7) is an appropriate clustering method to calculate dendrograms from these biological data (but note sources of error: ch. 8.2.3).

A simple distance measure is the portion of the differences between two sequences in relation to alignment length:

$$S = \frac{number\ of\ shared\ characters\ (nucleotides)}{N}$$

$$p = 1 - S$$

N: alignment length
p: visible distance
S: similarity

This is the visible **p-distance** (see also Hamming-distance, appendix 14.3.1). Another calculation for ***p***: if *N* is the length of two aligned sequences and *n* the number of substitutions (positions which do not show the same nucleotide in both sequences), then $p = n/N$. Example: when 10 positions of two sequences show two different nucleotides, we get p = 0.2. Note: theoretically the maximal evolutionary distance is 100 % (p = 1: no shared characters). The maximal visible distance of real data however is only about 75 %, because of four positions in an alignment of two sequences, one position has statistically identical nucleotides by chance alone.

The **patristic distance** (sometimes also called tree distance) is the number of character changes occurring on the branches that connect two terminal taxa in a given topology. This depends on the topology and can be obtained directly with parsimony methods, because these rely on counts of character changes (see F-ratio, ch. 14.10).

Distance: number of visible or estimated differences in a complex character (e.g., a gene) of two species or organisms in relation to the total number of identified detail homologies (e.g., number of alignment positions).

Visible genetic distance or p-distance: number of differences between sequences that can be directly counted, in proportion to alignment length.

Patristic distance: number of character state changes on the path between two terminal taxa along a given topology.

Evolutionary distance, d-distance: estimated number of substitutions which occurred in the course of evolution, counted on the path between two terminal taxa. This value is higher or equal to the visible distance.

Ultrametric genetic distance: genetic distance that is proportional to the real divergence time.

Divergence time: time elapsed (for clonal species, single organisms or clades) since two lineages leading to terminal taxa diverged from their last common ancestor, or (in the case of characters) since the replication in the last common ancestral organism that gave rise to separate character lineages.

Attention: usually, in dendrograms the branch lengths between nodes are drawn directly proportional to the estimated distances. The terms used in this context have to be distinguished clearly:

- Number of supporting positions on a branch: this is the estimated or visible number of substitutions which support a split in the dendrogram. Example: if two out of 100 nucleotides of a sequence are replaced on a branch between two nodes, the absolute number of supporting positions for this branch is 2; the frequency of the split, i.e. the proportion of positions representing the split in these 100 positions is 0.02.
- The p-distance between the two nodes is in this example 0.02.
- The substitution probability along the branch is 0.02 per position (uncorrected for multiple substitutions).
- Substitution rate: substitution probability per unit of time.
- Branch length: this can be defined arbitrarily. In distance and ML-methods it usually represents the substitution probability for the whole sequence in the time separating two nodes, and thus is the estimated number of substitution events. In parsimony methods the depicted branch length can symbolize the number of potential apomorphies, or the number of

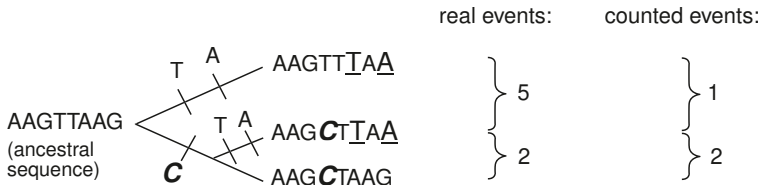


Fig. 164. Model for a true phylogeny with distortion of evolutionary distances by analogies (only 2 of 3 distances are shown) in a case with unequal substitution rates. Synapomorphies are shown in bold face, analogies are underlined. Concerning the terminology: note that analogies are not the same as homoplasies, see ch. 4.2.2.

character state changes in the most parsimonious topology, or some other support measure, or it even may have no relation at all to probabilities or support values.

8.2.3 Falsifying effects

The visible p-distance is of course influenced by each difference which exits between two sequences. The distance is increased by

- autapomorphies of a sequence,
- apomorphies of monophyla, when two sequences belong to different monophyla.

It is decreased by

- chance similarities and parallelisms of the two sequences,
- shared symplesiomorphies,
- shared synapomorphies.
- Furthermore, given a constant number of variable positions, the distance value is smaller when the number of invariable positions is larger.

A p-distance can only be interpreted as real evolutionary d-distance, when the following conditions are met:

- substitutions occur in all sequences on average with the same frequency,
- each variable position of a single sequence is substituted only once.

If these conditions do not apply to the alignment, distance corrections are necessary. The reasons for this necessity are found in the following considerations: when in a larger alignment two sequences of species that are not sister taxa evolve more rapidly than the other ones, they will show more chance similarities which reduce the distance between these sequences in comparison to the other ones (see Fig. 142). This has to be corrected. And: when several mutations occur at the same position of a single sequence, the visible distance is smaller than the real evolutionary distance (Fig. 165).

The longer the divergence time, the larger is the probability that more than one mutation occurs at a variable position (**multiple substitutions**) and that the substitution rates vary. Therefore

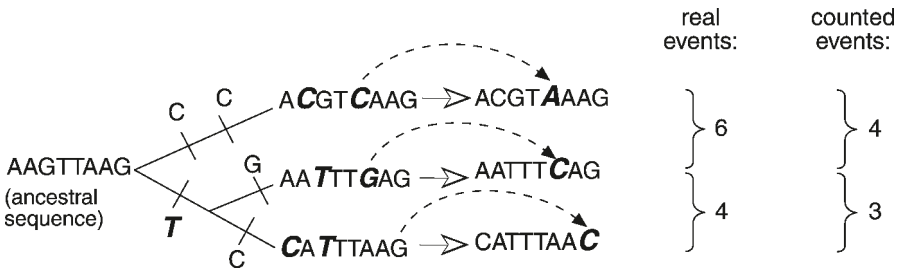


Fig. 165. Distortion of evolutionary distances by multiple substitutions (only 2 of 3 distances are shown): the real events are the number of substitutions which occurred on the path between the terminal sequences. Estimated events are the visible differences between terminal sequences. The nucleotides that were substituted last are shown in bold face.

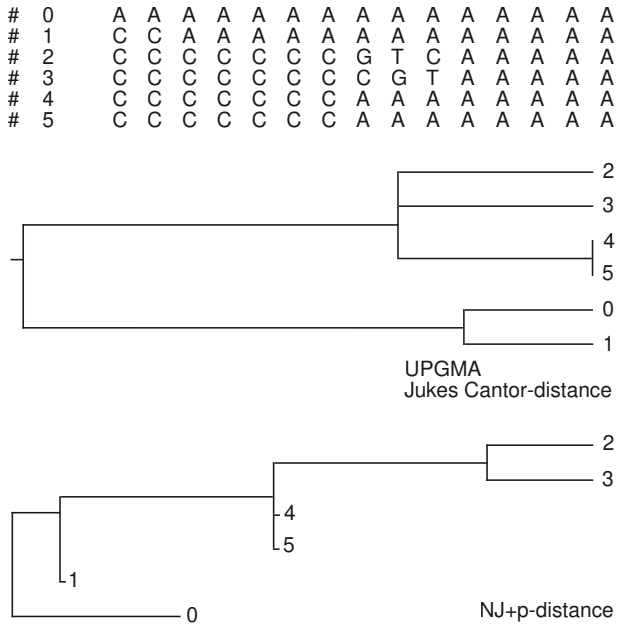


Fig. 166. Effect of trivial characters in distance analyses illustrated with a fictitious dataset. Sequence 0 is the outgroup, the sequences 2-5 show several synapomorphies (in each case a "C"), the sequences 2 and 3 each show three different autapomorphies (trivial, parsimony-uninformative characters) in the same positions. In the NJ-tree the sequences 2 and 3 are united as sister taxa, however, their similarity is based on symplesiomorphies and not on synapomorphies in comparison to sequences 4 and 5. The evolutionary process that unites them is the variability of the same 3 positions (positions 8-10). The correspondence between 4 and 5 in comparison to 2 and 3 is also based only on symplesiomorphies.

biological data are generally not perfectly additive and will show analogies. These deviations from character patterns in ideal additive data sometimes may have the consequence that relationships differing from the real phylogenetic ones are inferred.

With ideal distance corrections one obtains ultrametric distances which are proportional to the real divergence time (see ch. 14.3.4). In practice this is hardly ever achieved. In order to reduce the sources of error which exist when topologies are estimated from distance values, methods have to be used which do not require ultrametric distances and tolerate variations of substitution rates (e.g., neighbour joining algorithm, ch. 14.3.7). And corrections have to be introduced which compensate for the effect of multiple substitutions and of analogies (see below, ch. 8.2.6), so that the observed distances become additive. Distance methods will give you no indication of whether the correction has been successful or not. Therefore, as with every other method of phylogeny reconstruction, the plausibility of the result of a phylogenetic analysis has to be tested with information gained from other sources (s. ch. 10).

8.2.4 Effect of invariable positions, positions with different variability, alignment gaps

When invariable positions are inserted into an alignment, the distance values decrease, but the relations of the distance values between taxa stay the same. It is however a problem that many substitution models (see below) require the assumption that *all* positions of an alignment are equally variable, with the effect that the number of predicted substitutions may be higher than the real one. By eliminating invariable positions from an alignment it can be tested whether this changes the topology. The effect is often small, but it has to be tested empirically if this is a source for mistakes. Real data are even more complicated because the number of invariable positions may vary between taxa.

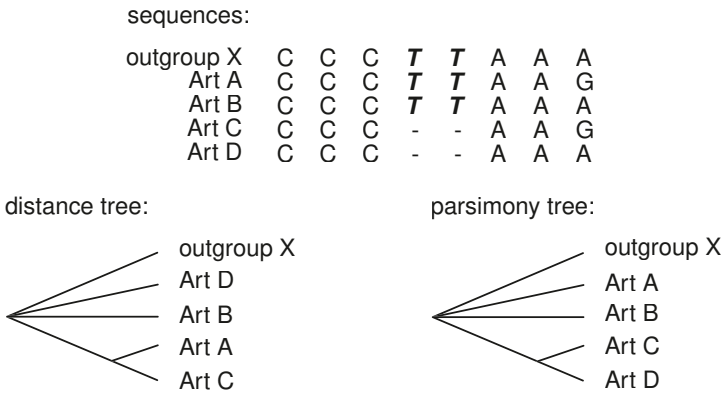


Fig. 167. Model of an alignment constructed to illustrate the problem arising when deletions and insertions are not considered as evolutionary events. Species X is the outgroup and has a plesiomorphic sequence. Species C and D have a common ancestor, in which a loss mutation occurred (deletion of TT). Species A and C show a convergence (G). Computing a distance tree (p-distance, neighbour-joining clustering method) the sequences A and C appear to be identical, because the positions with gaps were ignored. The parsimony method on the other hand finds the correct group {C+D} when the gap is coded as homology ("fifth base"), because in this case the number of potential synapomorphies (loss of two "T") is larger than in group {A+C}. This result requires the *a priori* hypothesis that the deletions are homologies.

When positions have a different variability they do not contribute to the estimation of divergence time to the same extent, because rapidly evolving positions get noisy earlier due to multiple substitutions (saturation effects: Fig. 43). Assuming for protein coding genes that synonymous substitutions occur more frequently than those which cause an exchange of amino acids (see ch. 2.7.2.4), a different rate can be used for types of codons (e.g., by differentiation of codon positions with 4, 2 or 0 possible synonymous nucleotide substitutions: Li et al. 1985). Another way to weigh substitutions consists of the translation of a DNA sequence into the corresponding amino acid sequence and to weigh the substitution of amino acids with values derived from empirical observations. Dayhoff et al. (1978) gained data on the relative frequency of amino acid substitutions which mirror the chemical properties of the amino acids, by comparing numerous proteins coded by nuclear genes (see ch. 5.2.2.10). The matrix of substitution probabilities can be used for distance corrections.

One can proceed in a similar way with rRNA-genes by empirically classifying positions of an alignment according to how many sequences show deviations in pairwise comparisons at a position. This percentage is compared with the distance of the two sequences to calculate an average value for the variability of each position (Van de Peer et al. 1993). After weighting the positions accordingly, a new distance matrix is calculated. However, it is doubtful whether these calculations will find the real evolutionary distance; the plausibility of the phylogeny is the only useful criterion for the evaluation of results.

During the development of distance methods the effect of **insertions** or **deletions** (indels) has been neglected so far. These become visible in alignments by the presence of alignment gaps. As the distance between a gap and a nucleotide (Fig. 167) is not defined in the available methods, there are only three alternatives to deal with indels:

- to ignore all positions containing gaps, which can imply a loss of information,
- to exclude the positions with gaps only in pairwise sequence comparisons,
- to treat gaps as "fifth nucleotide".

The second method has the consequence that in certain cases each sequence pair of a data matrix shows a different number of alignment positions (typical for rRNA-data). The example of Fig. 167 clarifies that under certain circumstances the evolutionary events cannot be considered with simple distance methods.

The definition of the alignment gap as "fifth nucleotide" produces mistakes, because this implies the assumption that the probability for gaps is the same as for nucleotides. One has to consider that the transformation of a nucleotide into another one (A→C) is a relative specific event, while the evolution of a gap (a deletion is a negative character (!); compare table in Fig. 102) can be with greater probability an analogy (the result of A→– and of C→– is the same).

When currently available substitution models are used for distance corrections (see below), positions with alignment gaps cannot be considered, because models for insertions and deletions have not been developed.

8.2.5 Effects of nucleotide frequencies

As DNA sequences are composed of only four characters, similarities between sequences very often originate from random matches of character states. Deviations of the nucleotide ratio from equal distribution (A:G:C:T = 1:1:1:1) increase the number of chance similarities. It sometimes happens that dichotomies in trees are determined by base frequencies (Steel et al. 1993). The effects of this unequal nucleotide distribution can be corrected with phylogenetic methods that use appropriate model parameters (see below). A further problem occurs when sequence regions consist for example only of A and T: such regions are not alignable and have to be eliminated from the dataset.

Nucleotide frequencies of single sequences or of groups of sequences can be calculated with several of the available computer programs (e.g., with MEGA, Kumar et al. 1993), the homogeneity of the base distribution in an alignment can be tested with the χ^2-test (say: chi-square), which also is implemented in available software (e.g., in PAUP). If some sequences appear as sister groups in a topology and at the same time show a significant bias in base frequencies, the grouping may be the result of chance similarities and therefore polyphyletic. In case the bias is found only within pyrimidines or purines it might help to code the sequences in the RY-alphabet.

χ^2-test of base composition homogeneity: to test if a sequence deviates from the expected average nucleotide frequency, the observed frequency n_{si} of sequence s and nucleotide i is compared with the expected number n_{ei} of nucleotides that should occur in this sequence when base composition is homogeneous in the alignment. The average frequency of "A" is the total number of "A" in the alignment divided by the total number of nucleotides (e.g., 3000/10 000 = 0.3), and the *expected number* in sequence s is the average multiplied with sequence length l_s of s (e.g., if $l_s = 2100$: $n_{ei} = 0.3 \cdot 2100 = 630$). The chi-square value is obtained with

$$\chi^2 = \frac{(n_{eA}-n_{sA})^2}{n_{eA}} + \frac{(n_{eG}-n_{sG})^2}{n_{eG}} + \frac{(n_{eC}-n_{sC})^2}{n_{eC}} + \frac{(n_{eT}-n_{sT})^2}{N_{eT}}$$

Look up this value in a Chi-square table assuming three degrees of freedom and a confidence limit of, for example, 0.1. If the calculated sum is larger than the value in the table, the null hypothesis is rejected and it must be assumed that the base frequency is significantly different from the average.

8.2.6 Distance corrections

Distance corrections are necessary to compensate distorting effects which cause a deviation of the visible p-distance from the real (evolutionary) d-distance. For this purpose one can refer, for example, to a correction curve (Fig. 168) from which a correcting parameter can be read which depends on the measured p-distance. The same goal is reached with a corresponding formula that represents the correcting parameter.

The simplest correction is the estimation of the expected number of multiple substitutions, assuming that the base frequency is 1:1:1:1 and is constant in time, and that only one constant substitution rate exists for all positions of the (correct) alignment (this is the Jukes-Cantor-model, compare appendix 14.1.1). Under these axiomatic conditions the **d-distance** can be calculated from the p-distance with the following formula (chapters 14.1.1 and 14.3.2):

$$d_{JC} = -\tfrac{3}{4} \ln (1 - \tfrac{4}{3} p)$$

The result is a new data matrix with corrected values (Fig. 162). This Jukes-Cantor-model can also be illustrated as a graph (Fig. 168) that shows the relation between p- and d-distances.

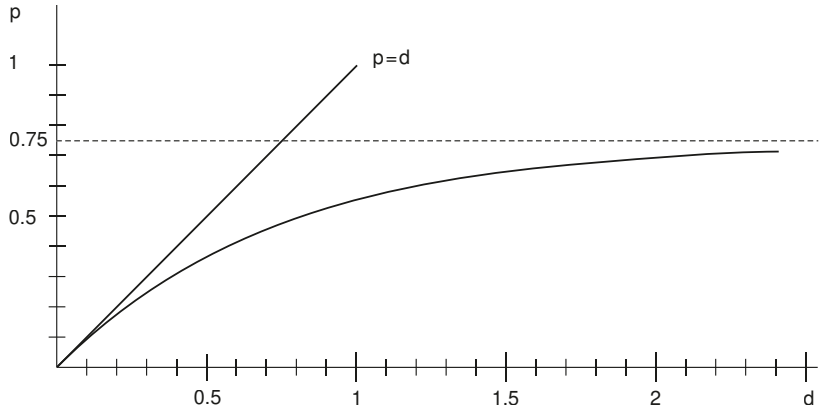


Fig. 168. Correction curve of the distance between two sequences for the Jukes-Cantor-model. The observed distance ***p*** is only with very small values (<0.1) approximately equal to the estimated evolutionary distance ***d***. The observed distance deviates from the diagonal, because multiple substitutions and random matches reduce the observed distance with increasing real (temporal) distance. The theoretical threshold of the p-distance is 0.75.

In the same way more complex methods can be used, such as counting the ratio of transversions to transitions and then assuming that for each of these substitutions an individual rate exists (Kimura-2-parameter model; appendix 14.1.3). Further models are listed in Fig. 159 (see also appendix 14.1 and Fig. 160). The models are usually integrated in computer programs designed to calculate distance dendrograms. These models are also important for maximum likelihood methods (chapters 8.3, 8.4, 14.6).

Take into account that absolute substitution rates cannot be estimated using models if no additional information is available. When it is stated that a constant substitution rate λ is used for the distance correction, this implies that only the substitutions which occurred since the divergence of two species during the unknown time ***t*** are counted or estimated from the given data and then it is assumed that this number corresponds to the product $2\lambda t$ (the evolutionary distance between two sequences; see Fig. 163). To estimate the absolute rate it is necessary to find a point on the tree for which a correctly dated fossil is known ("calibration of the molecular clock": see ch. 2.7.2.3).

The parameters used in models to transform visible distances are obtained by pairwise sequence comparison:

- **Count of visible distances:** compare two aligned sequences of length *N* and count the *n* sequence positions which do not show the same nucleotide in both sequences. The portion of these differences is the visible distance $p = n/N$ (see also the more general definition of the Hamming-distance; ch. 14.3.1).
- **Estimation of the relation of transitions to transversions**: compare two aligned sequences of length *N* and count the number of those sequence positions with nucleotides showing a transition difference (*T*s) and the number of nucleotides which show a transversion difference (*T*v). The respective portions are $P = Ts/N$ and $Q = Tv/N$. Note that $Ts + Tv = n$ and $Ts : Tv = P : Q$. Keep in mind that these are visible, uncorrected ratios!
- **Estimation of the base frequency (nucleotide frequency):** compare two aligned sequences of length *N* and count the number of nucleotides N_i (i = A,G,C or T) for each sequence. The average occurrence q_i of each nucleotide is calculated for both sequences with $q_i = (\sum Ni)/2N$. Furthermore the **portion of nucleotide pairs** which occur in the positions of two sequences of an alignment can be counted. For the pair A-T, N_{AT} is the number of positions with the pairing A-T or T-A, the portion is $x_{AT} = N_{AT}/N$. These calculations require the axiomatic assumption that the ancestral sequence had the average of the nucleotide frequencies of the daughter sequences and that invisible substitutions (multiple substitutions) did not occur or that they happened with the same frequency.

For cases with a large number of different rates which are specific for different positions of an alignment, the variability of rates can be described with a gamma distribution, which can be represented by a curve for the frequency of rates (see gamma distribution and other models in ch. 14.1.5). However, the application of a single gamma-correction to a distance matrix is based on the condition that the substitution rates remain constant with time and do not vary in different lineages. If there are reasons to assume that the substitution rate varies in time (***non-stationarity***) and when the base frequency is not homogeneous, a Log-Det distance correction is a recommended alternative (e.g., Lockhart et al. 1994, appendix 14.1.6).

Note that increasing the complexity of the substitution model the variance of distance estimates increases also (Zharkikh 1994, Li 1997).

Distances of protein-coding DNA sequences

When evaluating distances of protein-coding sequences, it can be considered that synonymous mutations are fixed in a population more often than non-synonymous ones. The reason for this phenomenon is probably in most cases the fact that mutations occur at the DNA level while selection acts on proteins. Single amino acids can be coded by 1, 2, or 4 nucleotide triplets. Correspondingly, probabilities for the occurrence of multiple substitution can be estimated on the basis of the relative frequency of synonymous and nonsynonymous differences and of transitions and transversions in the genetic code, depending on the kind of mutation (Li 1993; Pamilo & Bianchi 1993; see also Li 1997). These considerations are used for the correction of visible distances.

Working with amino acid sequences, a correction for multiple substitutions has to be performed for the same reasons mentioned for DNA-sequences (8.2.3). For a simple Poisson correction ($d = -\ln q$, when q is the portion of unchanged positions) one would have to presuppose that all amino acids have the same substitution probability. As this assumption is found to be wrong, weighting of individual substitutions can be performed with the help of a substitution matrix derived from empirical observations, which states how likely it is that a specific amino acid is replaced by another one (Dayhoff 1978; compare chapters 5.2.2.10, 14.11). A further assumption which is not correct is the independence of the substitution rate of the sequence position. In order to consider the variability between positions, Grishin (1995) suggests the following formula:

$$q = \frac{\ln(1+2d)}{2d}$$

where q is the portion of unchanged positions and d the estimated evolutionary distance. Applying this transformation to a complete distance matrix one has to assume that rates do not correlate with differences of life styles among organisms and corresponding differences of selection effects, that positions evolve independently, that the function of positions in the tertiary structure is not relevant for rate variations, and that rates remain constant in time.

Attention: the different variability of sequence positions in alignments of many sequences which indicates variations of selection pressure in different taxa is not considered in the distance methods described above. Only the average variability visible when comparing two sequences is counted.

8.2.7 Tree construction with distance data

Dichotomous dendrograms can be obtained from a distance matrix with the help of clustering methods (Fig. 162, see appendix 14.3.7). Network diagrams are constructed with split decomposition (appendix 14.4) and related methods. These methods rely on the fact that the species showing the smallest distances to each other should appear as closest relatives in a dendrogram. In principle the topology can be calculated "by hand" (see appendix 14.3.7), in practice one relies on fast computer programs.

UPGMA (= unweighted pair group method using arithmetic averages)

This clustering method mirrors the structure of distance data correctly only if distances are ultrametric (see also ch. 14.3.4). This assumption does not generally apply to biological data, wherefore

the neighbour-joining method should be preferred. The principle of the UPGMA method is explained in ch. 14.3.7.

Neighbour-joining

This clustering method does not require ultrametric distances and tolerates taxon specific deviations of substitution rates. The calculation is explained in appendix 14.3.7. Like all distance methods, neighbour-joining is sensitive for trivial characters (autapomorphies) which always modify distances (Fig. 166), and existing algorithms are susceptible to the order of taxa in the data matrix (the topology of the dendrogram can depend on the order of the taxa!).

Minimum evolution (ME)

This method is based on distance data, but in contrast to the previous ones it seeks to minimize the sum of branch lengths in the optimal topology. The method resembles therefore the maximum parsimony approach, however the tree length is computed from pairwise distances as in neighbour joining analyses (NJ) and the result depends on the model used to estimate the correct evolutionary distances. NJ and ME may generate different topologies (Nei & Kumar 2000, Takahashi & Nei 2000).

For an unrooted tree with n terminal sequences and $(2n–3)$ branches the sum of individual branch lengths e_i is

$$L = \sum_{i=1}^{2n-3} e_i$$

The tree with the lowest L is the minimum evolution tree. In practice, there exist several alternative least-squares methods to calculate the branch length e_i (Gascuel et al. 2001). Further details are explained in chapter 14.3.

Stemminess is the percentage of uncorrected minimum-evolution tree-distance attributed to internal branches (Lanyon 1988). It can be used as *a posteriori* measure for the signal to noise ratio in the data.

8.3 Maximum Likelihood: Estimation of the probability of events

In general, maximum likelihood methods estimate the likelihood that a hypothesis is correct when data and a model are given. The hypothesis is usually the tree, and the model parameters are *nuisance parameters* (e.g., branch length, base frequency, rate variation) that sometimes are not of interest but have to be specified. However, ML methods can also serve to estimate these parameters for the probability functions which describe the stochastic evolutionary process (assuming that character evolution *is* a stochastic process). Samples of species and of their characters serve as a starting point for the search of suitable parameters. In sequence analyses, the samples are the real sequences, and parameters that have to be estimated are substitution rates or the branch lengths of the optimal dendrogram that fits to the data. The values are optimized so that the sample (e.g., the sequences) can be the result of the estimated process with highest probability. Used for phylogenetics, these methods have the advantage that models can be precisely defined and applied for model-dependent analyses of the evolution of genes. The calculations are complex but can be accelerated with new methods (quartet-puzzling: see ch. 14.6; Bayesian analyses: see ch. 8.4).

The term "maximum likelihood" suggests that this is the (only) method that considers probabilities. This is not true. The evaluation of the similarity of complex characters, which is important for phylogenetic cladistics, is based on probability statements (ch. 5.1.1). The term should be considered as *terminus technicus*, it is a name for a specific method.

The application of process-oriented estimations of probabilities to phylogenetic analyses of DNA sequences is based on the work of Felsenstein (1981, 1993). The principle is likewise applicable to amino acid sequences (Adachi & Hasegawa

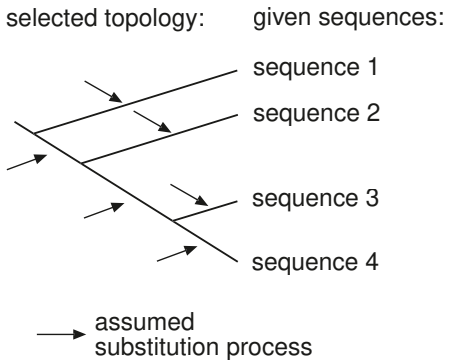


Fig. 169. The basis for maximum likelihood methods. The topology to be tested is chosen arbitrarily. The assumed substitution processes are described with models of sequence evolution. Different topologies are tested during the ML-analysis to see whether they can explain the evolution of the terminal sequences of the alignment with the given model. The maximum likelihood method yields comparable probabilities for alternative dendrograms.

1992). These maximum likelihood methods (short **ML-methods**) serve to estimate the probability that a phylogeny, as depicted in a given dendrogram, produces with a given evolutionary process the character distribution that has been observed in terminal taxa (Fig. 169). In order to enable this estimation, besides a data matrix, assumptions about evolutionary processes are also required, in the sense that the evolutionary rate (the probability for character state changes per unit of time) is assumed, or guessed, or estimated from patterns seen in the data. The assumed existence of specific evolutionary rates is described with a model of character evolution. Naturally, the result of an ML-analysis depends on the quality of the model. The latter has to allow the prediction that in a given period of time those substitutions which transformed an ancestral character into a character of an organism living today are possible with high probability. Apart from the model-specific assumptions it must also be assumed that the analysed sequence region is representative for the evolution of the species, that no sampling errors occurred, that the alignment contains the correct positional homology and that sequence evolution is a stochastic process.

The principle can be described as follows (for further details see appendix 14.6):

1) A dendrogram has to be selected for the studied species. In the course of the procedure all possible dendrograms can be tested to find the one fitting best to the data. The construction and search of dendrograms can be performed as in MP-methods (see appendix 14.2) or with the faster quartet-puzzling (ch. 14.6).
2) A model of sequence evolution has to be selected. Several parameter of the model (e.g., base frequency, Ts : Tv-ratio) can be estimated from the known sequences of the terminal taxa.
3) The probability is calculated that with the given topology and the presupposed substitution rates the terminal sequences could have evolved from a common ancestor sequence (further details in ch. 14.6).
4) The topology with the highest probability is selected from the possible alternatives.

In practice, the methods recommended by Felsenstein (1981, 1983), for example, have computation times so long that they can only be used with small datasets. As soon as a larger number of taxa have to be analysed (e.g., >25 species), a huge number of dendrograms have to be considered (Fig. 61) and the capacity of most computers is insufficient for this task. In these cases heuristic methods can be applied. A faster ML-calculation is possible with the **puzzle-method** (Strimmer & von Haeseler 1996), which estimates the most probable topology for four sequences at a time and assembles the complete topology from the comparisons of quartets (ch. 14.6). – Methodological improvements of algorithms and of computer techniques will increase the amount of data that can be dealt with; but in principle they cannot increase the reliability of an analysis as long as the axioms required by model-dependent methods remain untested.

Another promising variation of this principle is **Bayesian analysis**. With this method the *posterior probabilities* are estimated, in other words, the probabilities estimated *after* observing the evolution of a sequence along a topology when a model of sequence evolution is given. Using Bayesian algorithms one searches the tree or set of trees that maximize the probability of the tree for the given data and the selected substitution model (for more details see ch. 8.4). The basic model-dependent assumptions are the same as in ML-analyses.

Under specific conditions the result of an ML-analysis is identical to that of an MP-analysis (Tuffley & Steel 1997). This is theoretically true when each sequence position evolves independently, but not necessarily in the same way as the others, and when the substitution probability for a position along a branch of the topology is smaller than 0.5. This however, does not mean that MP- and ML-methods are interchangeable. ML-analyses are always studies of transformation processes and allow the consideration of very different substitution processes. One can also introduce process assumptions for parsimony methods, for example by weighting character transformations. However, the probabilities that certain substitution processes occur are not analysed, the aim is rather to test the compatibility of hypotheses of homology.

In principle, maximum likelihood methods can also be used for the analysis of morphological data (Lewis 2001). This requires definite assumptions about the process parameters driving character state changes. For example, one can assume that an average rate describes the expected number of changes along a branch and that the rate is symmetrical (it is the same for reversals). In theory, character-specific differences in rates (equivalent to positional rate heterogeneity) are adequately described with a discrete gamma distribution (ch. 14.1.5). However, the latter assumption requires that substitution rates do not change in the course of time for a single character. Since character states ("0" in character A and "0" in character B) are not comparable, the consideration of state frequencies is meaningless. A difference to parsimony analyses of morphological characters is that all characters, including those that show autapomorphies, are compared to estimate character variability in terminal branches, and that weights cannot be applied according to the complexity of the character change if state complexity is not considered by the model.

8.4 Bayesian phylogeny inference

Maximum likelihood methods as explained in the previous chapter allow to determine parameters of a model of character evolution prior to the selection of optimal trees. Even when estimating parameters from the data (base frequencies, rate heterogeneity of sites) one will assume that the prior experience on the quality of models will be useful to select among different trees (hypotheses). A Bayesian analysis injects information contained in the data based on the observation of how the data behave when constructing trees to improve the previous state of knowledge. In other words, this method is based on *posterior probabilities* of a hypothesis (Larget & Simon 1999, Huelsenbeck & Ronquist 2001 and references therein).

Posterior probabilities

Rolling a fair die the probability of observing one of six numbers is ⅙. Imagine you bought a new die and throw it twice. If you get a four and a six, the probability of observing these numbers under the hypothesis that the die is fair is

$$\frac{1}{6} \cdot \frac{1}{6} \cong 0.078.$$

Assume some dice are biased and that in these on average all numbers appear with equal frequency except the six, which occurs three times more often than any other number. This implies that on average one needs not 6 but 8 trials to get a "1", for example. The probability for all numbers is ⅛ except for the six, which has the probability ⅜. If you roll a four and a six with this die, the probability of observing these numbers is

$$\frac{1}{8} \cdot \frac{3}{8} \cong 0.047.$$

The probability of observing these numbers is reduced in this case (because the four is now rarer than in a fair die), while for a double six the probability increases to 0.14.

What is under these conditions the probability that a new die is biased when you observe a four and a six? Not knowing if a die is fair, but having the information that one out of ten dice shows the bias favouring the six, the prior probability $\boldsymbol{H_2}$ that a die is biased is $\boldsymbol{1/10}$, $\boldsymbol{H_1 = 9/10}$ is the probability that the die is fair. After rolling the die you will

have some information that helps to estimate if it is biased or not.

Bayesian probability estimation is based on the Bayes' law, which states that the *posterior probability* of a hypothesis H_2 is proportional to the *prior probability* of hypothesis H_1 multiplied by the likelihood of H_2 derived from the data collected. In our case, since the die was selected at random the probability $Pr(H_1)$ is 0.9, $Pr(H_2)$ is 0.1. From the available information about the properties of the dice we know that for the data ***D*** (observation of a four and a six) the probabilities are $Pr(D \mid H_1) = 1/6 \cdot 1/6 \cong 0.078$ (fair die) and $Pr(D \mid H_2) = 1/8 \cdot 3/8 \cong 0.047$ (biased die). Bayes' formula then yields for the probability of H_1 (the die is fair) in view of the observed data D:

$$Pr(H_1 \mid D) = \frac{\Pr(H_1) \cdot \Pr(D \mid H_1)}{\Pr(H_1) \cdot \Pr(D \mid H_1) + \Pr(H_2) \cdot \Pr(D \mid H_2)}$$

For the observed die the posterior probability that it is fair is then

$$Pr(H_1 \mid D) = \frac{0.9 \cdot 0.078}{0.9 \cdot 0.078 + 0.1 \cdot 0.047} = 0.94$$

After rolling the die (after making an experiment) we estimate that the probability of having bought a fair die increased from 0.9 to 0.94. $Pr(H_1 \mid D)$ is called a *conditional probability*, because it depends on the observed data.

The result of a step in a Bayesian analysis obviously depends on the *prior probability*. However, after a first experiment the observed data can be used as new starting point, the prior beliefs for the next experiment changed. As more and more data accumulate, the influence of the first priors decreases. The final results are likely to be insensitive to the priors.

In phylogeny inference, the unknown tree T_i is one of many possible topology states in tree space. When experiments are performed to obtain information about T_i (experiments are in this case sequencing projects to obtain alignments), the experiments are designed so that the observations are distributed according to some probability distribution which has T_i as an unknown parameter. If a single result of such an experiment is denoted X, the **posterior distribution** of T_i given X is denoted $Pr(T_i \mid X)$ and is defined as the conditional distribution of T_i given the sample observation X .

Markov chains and Monte Carlo algorithms

To find the optimal tree we have to sample tree space (as in parsimony and in other likelihood methods), which in Bayesian inference is restricted to the topologies that are contained in the posterior distribution. Bayesian analyses uses stochastic simulation to obtain samples from the posterior distribution. To construct topologies and select these samples, Markov Chain Monte Carlo (MCMC) algorithms proved to be computationally efficient.

Markov models* are based on the assumption that in a stochastic process an expected event depends only on the actual state, not on earlier ones. **Monte Carlo** simulations (named after the gambling casino in Monaco) are simulations in which specific assumptions on the model of sequence evolution are presupposed and single events are independent of the previous ones. The selection of single events occurs at random but considering their model-dependent probability. This is equivalent to rolling dice. A **Markov chain** is in our case a series of trees and model parameters obtained proposing a new hypothesis by modification of a previous one. For a stochastic transformation of one tree into another one with a Markov chain Monte Carlo (**MCMC**) simulation an initial tree is needed and a matrix defining the probabilities for the transformation.

All inference is based on the posterior distribution $Pr(T_i \mid X)$:

The hypothesis that will be tested is the tree $\boldsymbol{T_i}$ (which in this case represents a topology with branch lengths and model parameters) and the posterior probability is found by sampling from the tree space (from the posterior probability distribution). The prior probability for a single topology is $1/\boldsymbol{B}_{(s)}$ when $\boldsymbol{B}_{(s)}$ is the maximum number of trees that can be constructed from s species. This means of course, that at the beginning each topology is considered to have the same probability. However, we want to know the posterior

* named in honor of Andrei Andreevich Markov (1856–1922).

probability for a tree when the alignment (or matrix) $\boldsymbol{X}$ is given. Therefore, for $Pr(H_1)$ we can write now $Pr(T_i)$, and for $Pr(D \mid H_1)$ we write $Pr(X \mid T_i)$, which is the likelihood of the *i*th tree when the alignment *X* are the observed data. The denominator is composed of the sum of all $Pr(T_j) \cdot (X \mid T_j)$ for all topologies T_j (from topology number 1 to number $\boldsymbol{B_{(s)}}$) that can be constructed for *s* species:

$$Pr(T_i \mid X) = \frac{\Pr(T_i) \cdot \Pr(X \mid T_i)}{\sum_{j=1}^{B(s)} \Pr(Tj) \cdot \Pr(X \mid Tj)}$$

As in maximum likelihood analyses (previous chapter), each position of the alignment is examined to estimate the probability of observing the data assuming the given topology and a substitution model. This probability for the *i*th alignment position is the sum over all possible assignments of nucleotides to internal nodes of the given topology considering the substitution probabilities for each branch of the tree with the given substitution model (see ch. 14.6). The likelihood of a topology and a given alignment is obtained by multiplication of the likelihoods of each sequence position.

In Bayesian analyses the topologies and the model parameters are modified in a chain of steps to find the values that maximize the likelihood function.

Computing the denominator in the formula for $Pr(T_i \mid X)$ is infeasible for most datasets, because all possible topologies have to be considered. Since it is not possible to calculate the posterior probability analytically, an approximation is possible by sampling trees and the other model parameters from the posterior probability distribution. The procedure is the following:

A Markov chain is started with a T_i selected at random or from some previous analysis and a new T_i' is proposed. The modifications consider the probability for changes implied by model parameters. The so-called Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970) uses posterior probabilities to decide if the new tree T_i' is accepted in the chain. The new hypothesis is accepted with the probability

$$R = \min\left(1, \frac{\Pr(X \mid T_i')}{\Pr(X \mid T_i)} \cdot \frac{\Pr(T_i')}{\Pr(T_i)} \cdot \frac{\Pr(T_i \mid T_i')}{\Pr(T_i' \mid T_i)}\right)$$

This formula contains a multiplication of the likelihood ratio, the prior ratio and the proposal ratio. If the new hypothesis is not accepted, the original T_i constitutes the next sample. Essentially, the new tree T_i' is accepted if the posterior probability $Pr'(T_i' \mid X)$ is larger than $Pr(T_i \mid X)$. Uphill steps are always accepted, slightly downhill steps are usually accepted, but large downhill steps are almost never accepted. The effect is that the chain moves uphill and when reaching the plateau of the hill the probability values sway around the maximum of the most probable tree. This process of proposing new trees is repeated many thousand times. The frequency of occurrence of a single tree in this sample is an approximation of the **posterior probability of the tree**.

Running several Markov chains in parallel and swapping states between two chains, a chain trapped in a local optimum can escape to continue with a better tree (Huelsenbeck & Ronquist 2001). The so-called *cold chain* is the one that counts (the one that is sampled), the other ones are *heated chains* that act as scouts and test shorter and longer steps in the modification of topology and model parameters to find a way to escape a local optimum. This procedure is called **Metropolis-coupled MCMC**.

A 95 % **credibility interval** can be defined starting with the most probable tree ("on top of the hill") and then adding trees in order of decreasing probability until the cumulative probability is 0.95.

Each step in the modification of a chain is also called a generation. The Markov chain will be run many thousand generations (e.g., 100,000) and trees are sampled at regular intervals (e.g., every 100 generations). Starting with a random tree it takes some generations until the likelihood values appearing in the chain reach apparent stationarity. Therefore, the first generations are discarded (" burn in" of the chain). In practice, one can run, for example, four parallel chains and sum the natural logs of the likelihoods in each chain. This sum should converge upon a stable value before the sampling of trees from the chains begins. After stopping the chains a consensus tree can be constructed from the sampled trees.

The posterior **probability of a single clade** is the sum of the posterior probabilities of all trees that

contain that clade (or the number of times the clade occurs in the MCMC sample divided by the total number of sampled trees). Thus confidence values for branches are obtained that are comparable to bootstrap values. However, differences in bootstrap values are based on differences in the number of split-supporting characters and the effect and number of contradicting characters using a constant substitution model and a resampled dataset, while Bayesian probabilities are based on differences in the fit between a model and a dataset using modifications of models and a constant dataset.

And, in comparison with bootstrapping using the maximum likelihood procedure the Bayesian analysis is much faster. The cause for this difference is the time required by ML-methods to compute a single tree and the fact that during bootstrapping each tree is optimized separately while in a Bayesian analysis every tree in a chain is derived from a previous optimization.

To summarize the results one can depict the topology with the maximum posterior probability or a consensus tree and show the posterior probabilities of clades in a similar way as done with the results of a bootstrap analysis.

There exist several applications of this method. One can compare the posterior probabilities obtained for different datasets, but also the values for model parameters obtained for different partitions of an alignment to find out if a single model explains the evolution of a gene. Comparing the likelihood values of two models one can chose the model that permits a better explanation of the data (Huelsenbeck et al. 2001). Imposing a constraint on a node that defines the composition of a monophylum one can get estimates for the ground pattern sequence of this node (Hall 2001).

Attention: Despite its advantages, one should not forget that the Bayesian analysis in the version discussed here is based on the maximum likelihood approach and requires a model (e.g., the HKY85 model, see Fig. 159) and all the assumptions implied with the selected model. Model misfit will cause unjustified confidence in the result.

8.5 Hendy-Penny spectral analysis

The use of the Hadamard-conjugation (appendix 14.7) for the analysis of DNA sequences was developed by Hendy & Penny (1993) to estimate the phylogenetic information present in an alignment within the scope of **spectral analysis**. The spectrum (Fig. 170) is a histogram. In contrast to the one in Fig. 153, it does not show for each split of a dataset the observed number of supporting positions but illustrates branch lengths, or the estimated portion of supporting positions in which a taxon or species group differs from others.

The method differs from the computation of *spectra of supporting positions* (ch. 6.5.1) because (a) noise is not studied phenomenologically (noisy positions are not considered to belong to the set of supporting positions) and instead a correction is performed with the *help of models* of sequence evolution, and (b) each possible split of a set of sequences is considered using Hadamard conjugation instead of only those that are really represented in the data, wherefore this type of analysis is computationally very expensive. The following steps are necessary:

- For each sequence position of a given alignment all species groups showing the same nucleotide are recorded. Each group forms a split (species with nucleotide *i* vs. species without nucleotide *i*). Theoretically, for each position a maximum of four splits can be recognized (one split for each nucleotide). Simple algorithms, however, may require binary characters (e.g., DNA sequences in the R-Y-alphabet).
- The **frequency of a split** is the sum (number) of the sequence positions containing this specific split. The frequencies of all splits present in a dataset compose the *sequence spectrum* ("crude spectrum") used in further calculations.
- With the help of Hadamard conjugation a new "***r***"-vector can be calculated from the

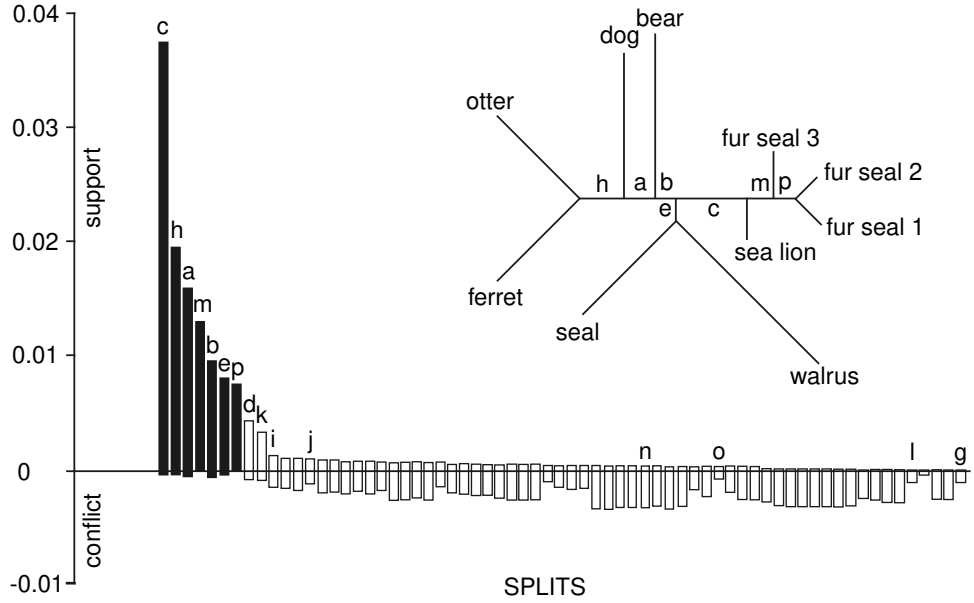


Fig. 170. Example of a Hendy-Penny-spectrum ("Lento diagram" from Lento et al. 1995: combined cytochrome b and 12S rRNA data). Splits are sorted according to their support, with the best splits at the left side of the spectrum. Correction of the distance between groups with the log-det transformation (ch. 14.1.6). The diagram shows the support of splits (above the x-axis) and the conflict (standardized support of incompatible splits below the x-axis, see text.). Columns for those non-trivial splits contained in the optimal topology are filled in black. It can be seen that the best splits are mutually compatible and thus fit to a single topology, and they show little conflict. In a corresponding diagram without log-det transformation the compatible signals are more scattered over the spectrum and conflict values are higher.

sequence spectrum, which contains all distances between groups of terminal taxa ("generalized observed (uncorrected) distances"; the distances are called "generalized" because taxa are not compared pairwise, but in bipartitions of the whole dataset).

– Generalized uncorrected distances observed for each split are transformed for the same purpose as in distance methods with a model of sequence evolution into another value which considers the number of estimated multiple substitutions ("***rho***"-vector with "generalized corrected distances").
– The vector of corrected generalized distances between the groups of all splits can be transformed into a "***gamma***-vector" of branch lengths of a topology (a tree spectrum) using the Hadamard matrix. For ideal data (without noise) the vector has as many entries >0 as the number of branches of a tree constructed from the data. These branch lengths are already corrected for multiple substitutions (due to the corrections in the previous step) and represent evolutionary distances. The *uncorrected* observed distances of a topology can be regained with the help of a model of sequence evolution. The method originally introduced by Hendy & Penny uses the Jukes-Cantor model.
– The number of conflicts for each split in a selected topology is recorded. A conflict is an incompatible split in a sequence position (for the term incompatibility see Fig. 55). For representation in the histogram this value is multiplied with a factor consisting of the overall ratio of supporting to incompatible values of the dataset (sum of all supporting values / sum of all conflicting values). Support and conflict of each split can be visualized in a single spectrum (Fig. 170).

The rho-vector is suitable as a starting point for the calculation of distance trees and the gamma vector can be used for MP- and compatibility methods. A detailed description is given in the appendix (ch. 14.7).

8.6 The role of simulations

Simulations are meant to imitate real processes and are used when it is impossible to deduce the total number of solutions due to the large number of variables. Simulations have been performed frequently to test empirically which method of tree construction and which type of data produce reliable results. To do so, data have to be produced which simulate the effect of an evolutionary process, starting from an ancestor. DNA sequences are well suited for simulations because they contain only four types of characters and the consequences of mutations can easily be constructed in descendant sequences. For the observance of a realistic speed of evolutionary changes, empirically observed mutation or substitution rates are used for models of sequence evolution. Using computer programs one can, for example, evolve a random sequence along a given topology. Monte Carlo simulation is able to estimate the likelihood of getting a specific result by running hundreds or even thousands of "trials". Either the phylogenetic tree and its branch support values obtained from artificially produced sequences or the structure of an alignment obtained after evolution along a topology are recorded, so that these results can be compared with those obtained from real data.

Simulations are useful, for example, to show if the model parameters (topology and substitution rates) assumed for a maximum likelihood analysis can yield the result obtained from real data.

In practice, however, simple simulations are not a perfect choice to test the reliability of computer programs. Often those tree reconstruction methods yield the best results in simulations which apply the same model of sequence evolution also used to generate the sequence alignment. The conclusion that one has found the best method is circular in this case.

Also, models used for simulations are probably in many cases too simple. The comparison of branch lengths, for example, measured as substitutions per site, are not realistic when the sites are assumed to have the same rate. While sequences in a real alignment may show a relative low number of substitutions per site, this can result from the presence of a large number of invariable sites, while the positions that are free to vary may be saturated. So, regarding the variable sites, the sequences may be evolving very fast, while using the complete sequence a short branch estimation results that suggests no saturation (the **hidden long branch** is not discovered). Simulations without consideration of this phenomenon will suggest an illusive confidence in the results of analyses of real data.

Nevertheless some relevant information has been obtained with simulations: it was possible to show that for the use of distance methods there exist no universal rules for the selection of weights and substitution parameters. Correction of p-distances produce a large variability, wherefore corrections are not efficient for large real substitution rates (Schöniger & von Haeseler 1993). Lower weighting of frequent events (transitions, substitutions of third codon positions) improves the result, because positions which are less noisy and more informative get a higher weight. With low substitution rates, better results were obtained by weighting of substitutions per sequence position in an alignment, with higher rates by weighting of substitutions per sequence pair. Furthermore, in distance methods the variability of transformed distances is larger when the divergence times are larger, and those methods which consider more parameters sometimes enable better reconstructions in some simulations (compare Tajima & Nei 1984, Zharkikh 1994). This, however, does not hold anymore when the substitutions approach saturation and multiple substitutions are common: with increasing distances the refinement of models shows little effect (Rodriguez et al. 1990). For the application of simulations in the framework of parametric bootstrapping see ch. 6.1.9.2.

9. Sources of error

In the preceding chapters several errors that are specific for single methods were discussed. They can be reduced to a few fundamental problems. When these are known, it is relatively simple to find possible sources of error in individual cases. One reason for the lack of consistency of methods of phylogeny inference are violations against the implied axiomatic assumptions (remember: methods are consistent when their ability to recover the correct phylogeny increases with the amount of data). Wrong axiomatic assumptions about evolutionary processes, for example, are mistakes which cannot be tested with the respective tree constructing methods.

9.1 Overview of common sources of error

In principle, in the course of each phylogenetic analysis the following mistakes can occur; these are based on violations against general axiomatic assumptions needed by tree reconstruction methods:

The sample of individuals is not representative. Some individuals can possess characters which are not representative of the species. These are new characters (e.g., of local races or individual mutations) not belonging to the ground pattern of a species. Possible consequences: individuals are placed in the dendrogram outside of the species to which they belong.

The species sample is not representative. No tree reconstruction method can test whether species sampling is sufficient to reconstruct the phylogeny of a species rich taxon. A frequent source of error, especially in molecular systematics, are symplesiomorphies which support paraphyletic groups (ch. 6.3.3). Furthermore, the danger exists that analogies support polyphyletic groups when the genetic distances between the selected species are too large (ch. 6.2.2).

Terminal taxa are not monophyletic. The relationships to other taxa cannot be recovered correctly when terminal taxa are not monophyletic. For example, apparent outgroups could in reality be part of a terminal taxon of the ingroup (Fig. 137).

The sample of characters is not representative. As characters always represent hypotheses of homology, there exist "good" and "bad" characters containing traces of phylogeny with high or little probability (ch. 5). Therefore, character selection influences the result of phylogenetic analyses. No phylogenetic algorithm can test whether there are better characters besides the information contained in the data matrix. Specific errors in selecting characters can occur that are independent of whether these are morphological, molecular, or ethological characters:

A) The selected character is not a **character of the ground pattern** of the terminal taxon but an autapomorphy of a subsets of species of the terminal taxon. The error occurs more often when no ground patterns were reconstructed for terminal taxa. Species-specific sequences which were obtained from single individuals have to represent the features shared by all individuals of a species, and therefore should, if possible, show few population-specific or individual autapomorphies. Individual sequences which represent large supraspecific taxa should be reconstructed ground pattern sequences of these taxa (this approach is still not common practice). In the same way morphological characters must be judged to be ground pattern characters of the represented taxa.

B) The characters are not homologies and do not represent real objects or properties of nature, but are products of **fantasy**, **superficial similarities**, the result of erroneous perception or laboratory mistakes. Therefore non-monophyletic groups are obtained.

C) **Frame homologies** of terminal taxa are not homologized correctly. "To homologize correctly" means that only such properties or structures are combined in discrete frame characters for which also the probability of homology of the frame has been estimated. This includes the homologization obtained aligning sequences. Genes which are compared for phylogeny inference have to be orthologous. In comparative morphology, the frame homologies (e.g., "eye") have to be homologous with high probability to allow the homologization of character states (of detail homologies) like "eye with lens" and "eye without lens". Character states of eyes of scallops should not be coded in the same column as the eyes of vertebrates or of cephalopods. These eyes evolved independently in different parts of the body (mantle rim, mollusc head, vertebrate head).

D) The **probability of homology** of the selected characters is very low and therefore random groupings of taxa are obtained.

E) The probability of homology of **unweighted characters** may be very different, important characters are not distinguished from "unimportant" ones (see ch. 5). The consequences: real synapomorphies contribute little to the reconstruction of the phylogenetic tree, which therefore can become incorrect in many parts.

F) The characters are **not weighted correctly**. Weighting has to refer to (a) the probability of cognition or (b) the probability of events, and is a statement on (a) how probable it is that a homology can be recognized correctly without assumptions on processes, or (b) that a character state could have evolved from another state by a specific process. Morphological characters have to be compared phenomenologically (ch. 5.2.1). For sequences, the positional homology (ch. 5.2.2) and, in the ideal case, also the relative probability of the homology of character states (ch. 5.2.2.2) have to be determined when the intention is to do a phenomenological analysis. When modelling methods are used for sequences, the substitution models have to be as close to reality as possible.

Characters are non-independent. It may be that only one event occurred but due to pleiotropic effects more than one character changes. The result is an overestimation of the weight of this event.

The selected sequences are not informative. This is a special case of insufficient character sampling which has to be stressed because sequences are often used without control of their phylogenetic information content. Cause for insufficient information content can be (a) a too rapid evolution of the sequence or a long time of isolated evolution of single branches, so that the "homology signals" become completely noisy due to substitution of apomorphies, or (b) sequence evolution is too slow or stemlines are too short so that no synapomorphies evolved and random similarities determine the topology.

Reconstruction of phylogeny without evaluation of character quality (phenetic analysis). Character analyses are essential for phylogenetic cladistics to determine *a priori* character polarity and probability of homology, the latter is used for differential weighting. The philosophy behind this statement is that the estimation of data quality and the use of data for inferences are two different and independent steps. Modelling methods imply a character analysis in which the probability of character transformation is estimated. If the model is adequate, one also gets a type of weighting. Abstention from these analyses can have the consequence that chance similarities and plesiomorphies influence the topology of a phylogenetic tree.

Algorithms imply unrealistic axioms. Often phylogenetic algorithms require assumptions which have the function of axioms. These can be assumptions about the process of evolution or about properties of the data (Fig. 159). Possible consequences: the reconstruction is not an image of the historical processes and the topology is not correct despite satisfactory tree statistics. Some typical axiomatic assumptions:

A) **Distance methods:** the analysed sequence region does not contain sampling errors (is representative for the average of stemline substitutions). The evolution of this sequence is a stochastic process (drastic episodic rate changes do not occur). The estimated distances cor-

respond to the real evolutionary distances, or rather the assumption implied by the models used are realistic and allow the reconstruction of correct evolutionary distances. Most substitutions in an alignment are not "saturated" (synapomorphies did not erode).

B) **UPGMA clustering methods:** in addition to the assumption of all distance methods it is required that substitution rates are equal for all stem lineages; distances are ultrametric.

C) **Parsimony methods:** the estimated probability of homology of unweighted characters is equal for all characters, or character weights correspond to relative differences of estimated probabilities of homology. This also includes assumptions about the reversibility of character states that can be incorrect (ch. 6.1.2).

D) **Maximum likelihood methods:** the assumptions of the model selected to represent the process of sequence evolution are realistic. The analysed sequence regions do not contain sampling errors and are representative for the average of stemline-substitutions. Sequence evolution is a stochastic process and the model is valid for all parts of the tree and for all sequence regions.

Data contain laboratory artefacts: there are many sources for errors especially when molecular data are used. Some examples: amplification of contaminations via cloning or PCR, misidentification of specimens, specific laboratory mutations, systematic misreadings, sample crossover, wrong assemblage of gene fragments (not only within data from a single species, but also confusing data from different species), miscopying in a data table.

9.2 Criteria for the evaluation of the quality of datasets

Many of the preceding chapters are dedicated to the question of the assessment of data quality. It has been explained why only apomorphies can provide evidence for relationships (ch. 1.3.7, 3.2.3), that apomorphies are hypotheses of homology, and that it has to be estimated for these hypotheses whether they are well supported by the available evidence (ch. 4 and 5). Furthermore, it is important that the species sample is representative for the taxa of interest (ground patterns: ch. 5.3.2) and for the analysed part of the tree of life (polyphyly due to long branches: ch. 6.3.2; plesiomorphy trap: ch. 6.3.3). We have to distinguish *a priori* criteria which are used to estimate data quality independently of the result of a data analysis, and *a posteriori* criteria that measure the fit between data and topology.

***A priori* criteria** (before tree construction):

The quality of a dataset is improved,

- the more careful the probability of homology of characters (positional homology, frame homology) and of character states have been evaluated to select a weighting scheme or to exclude characters;
- the more terminal taxa were examined for the occurrence of character states;
- the more details of ground patterns of terminal taxa were reconstructed by phylogenetic character analysis;
- the more characters of high probability of homology were used or the more phylogenetic information (apomorphic detail homologies) was identified in morphological characters or in sequence alignments and the better the signal to noise ratio;
- working with sequences: the more species representing the genetic diversity of a taxon are considered and the closer basal species are to the common ancestor to increase the probability that ground pattern characters (plesiomorphies within the ingroup) are reconstructed correctly;
- the more careful "long branches" were identified and avoided by exclusion of rapidly evolving species or by inclusion of additional taxa that divide long edges into shorter inner branches.
- The higher the signal to noise ratio, the more reliable are the data (ch. 6.5).

In this list the criteria of absence of conflict in a phylogenetic hypothesis, or satisfying tree statistics of a cladistic analysis, or positive results of permutation tests and so on (ch. 6.1.9) are not

mentioned, because these are *a posteriori* criteria that describe the fit between trees and datasets but are not topology-independent indicators of data quality.

On the evaluation of characters. According to the considerations introduced in chapters 4.1 and 5, the information content or quality of characters depends on the number of informative details and their probability of homology. In order to estimate character quality independently of any further assumptions about the course of evolution, a *phenomenological character analysis* has to be performed (see. ch. 5). For sequences, careful alignments are necessary (ch. 5.2.2) and noisy regions have to be identified to exclude or downweigh characters of unreliable positional or character state homology. For the evaluation of the relative probability of homology of nucleotides (representing putative homologous character states) spectral analyses are recommended (ch. 6.5).

On the selection of species. It is of course easy to detect whether datasets differ in the number of species considered. A more difficult problem is to find out if the selected species are representative for a taxon. Since the reconstruction of phylogeny depends on the correct **inference of ground patterns** to represent larger terminal taxa or larger sections of a tree(ch. 5.3.2), it is essential that those species are sampled which probably do not deviate much from the ground pattern of the monophyla they should represent. These are usually species which morphologically look primitive or little derived, or which are similar to related outgroup taxa. Obviously, the famous model organism *Drosophila melanogaster* (Diptera) is not a good choice to represent insects, and other genetically well studied species (*Caenorhabditis elegans* for Nematoda, *Homo sapiens* for Mammalia) certainly were not selected having their relevance for phylogenetic studies in mind. When it is not known which species are particularly primitive, one can try to include in the analysis as many species as possible. Ideally, we would like to consider all described species of a taxon. Unfortunately, at the moment this is only possible for some morphological characters in well-studied taxa. Furthermore, **long stem lineages** can be avoided especially for sequence analyses by consideration of many species, with the effect that the danger of appearance of false monophyla due to symplesiomorphies (ch. 6.3.3) and analogies (ch. 6.3.2) is reduced. For this purpose those taxa should be chosen as outgroups for which one can assume that they show the plesiomorphic state for characters that are modified in the ingroup. Using morphological characters one can often consider at least for character analyses (ch. 5), all species of the outgroup (all organisms not belonging to the ingroup).

Examples for ***a posteriori* criteria** (comparison of data and topologies):

- number of branch-supporting discrete characters
- bootstrap proportions and jacknife percentages (ch. 6.1.9.2)
- decay index (= Bremer's index; ch. 6.1.9.2)
- stemminess (ch. 8.2.7)
- likelihood value (ch. 8.3, 10.2)
- posterior probability (ch. 8.4)

When the results of a phylogenetic analysis are not plausible (plausibility: see below, ch. 10) or contradicting other data of high quality, then doubts are justified concerning (a) the quality of the data and (b) the method of tree reconstruction.

10. Comparison of dendrograms and plausibility tests

10.1 Plausibility

A dendrogram is plausible when it is compatible with *data that were not used* for its construction and that can be explained with a phylogenetic hypothesis. Those criteria derived from analytical methods that compare data with topologies (bootstrap-test, branch lengths, number of apomorphies, rate tests, etc., see preceding chapters) do not belong to this category, because they fall back on the same methods and/or data which had been used for the generation of the dendrogram (see testing of hypotheses in chapter 1.4.2). Additional data suitable for the test of plausibility are

- dendrograms obtained with other characters and other methods of phylogeny inference,
- historical-biogeographic patterns,
- age and characters of fossils,
- considerations on the adaptive value of evolutionary novelties and on the plausibility of character evolution.

Comparison of different datasets

When the same topology is reconstructed with independent analyses, the probability that congruence is not due to chance is large, especially when many species have been considered (compare number of alternative topologies, ch. 3.4). But it has to be taken into account that a non-accidental congruence of topologies can nevertheless be misleading when the same selection of taxa produces paraphyla by plesiomorphies. The plesiomorphy trap will be effective for many characters or genes (ch. 6.3.3). Another source of error could be a bias in base composition occurring in several genes, or a systematic error caused by an algorithm when the same method is used for independent analyses of several genes. Noteworthy are congruencies which were obtained by independent analysis of morphological and molecular characters (Fig. 171).

Equally interesting are also contradictions between morphological and molecular analyses: it would be relevant to test which of the contradict-

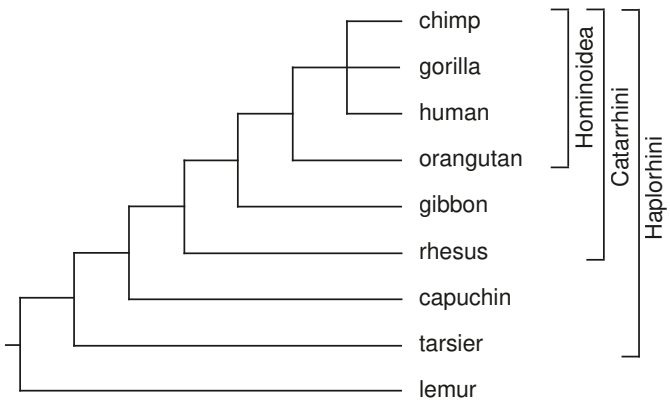


Fig. 171. Correspondence between molecular and morphological characters: phylogenetic tree of primates based on globulin gene sequences (after Goodman et al. 1994). The monophyly of taxa is supported among others by the following morphological characters: Haplorhini: similar derived features of the anatomy of the auditory meatus, of the tympanic region, the dentition, the placentation. Catarrhini: e.g., only 2 premolars per half denture, molar 2 and 3 with hypoconulid, nostrils close to each other, thumbs completely opposable. Hominoidea: e.g., size of brain, reduction of the caudal vertebrae (compare Starck 1995).

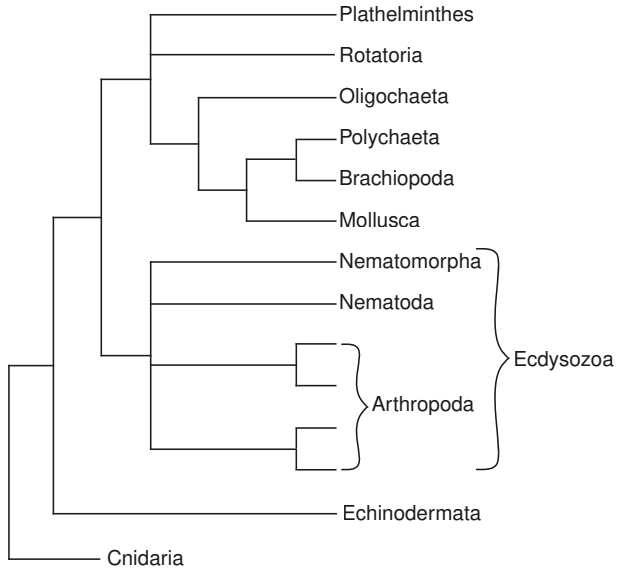


Fig. 172. Example for contradictions between molecular and morphological analyses. The Ecdysozoa hypothesis was originally based on an 18S rDNA alignment (Aguinaldo et al. 1997) and is not in accordance with morphological data (Fig. 174). A spectral analysis of the original alignment shows that the signal in favour of the Ecdysozoa is not better than the background noise (Wägele et al. 1999, see also Fig. 149), but compatible with a most parsimonious solution, while apomorphies for arthropods and within arthropods are rare and not affecting the topology. In view of the weak signal (Fig. 173) the alternative hypotheses (compare with Fig. 174) should not be dismissed.

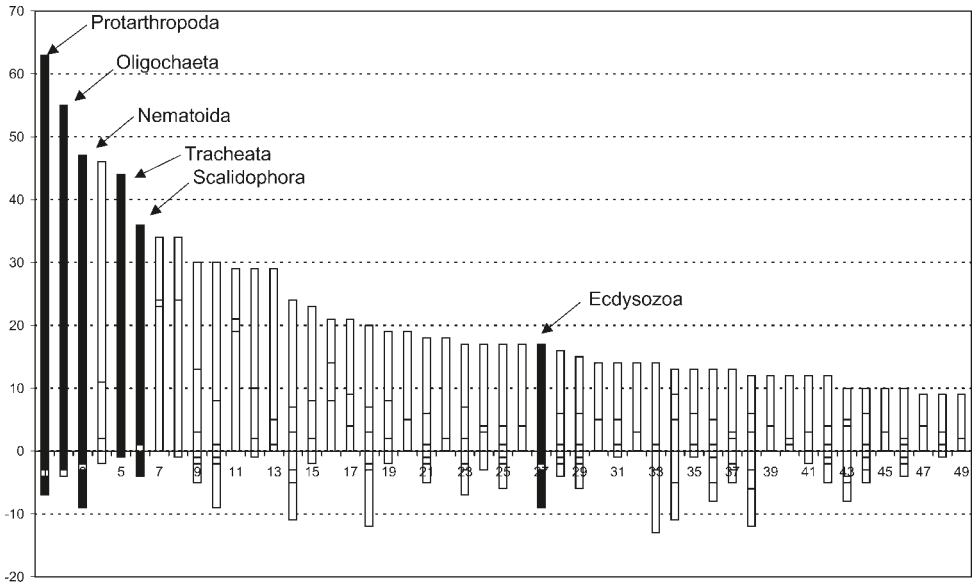


Fig. 173. Spectrum of supporting positions for the original alignment used to postulate the Ecdysozoa hypothesis (Aguinaldo et al. 1997). Only the best supported splits are shown. Black columns indicate the support for splits that appear in the optimal tree, white columns are support for splits incompatible with the optimal tree and indicate the level of background noise. The Ecdysozoa split is compatible with the optimal tree but the support is not better than the background noise (from Wägele et al. 1999).

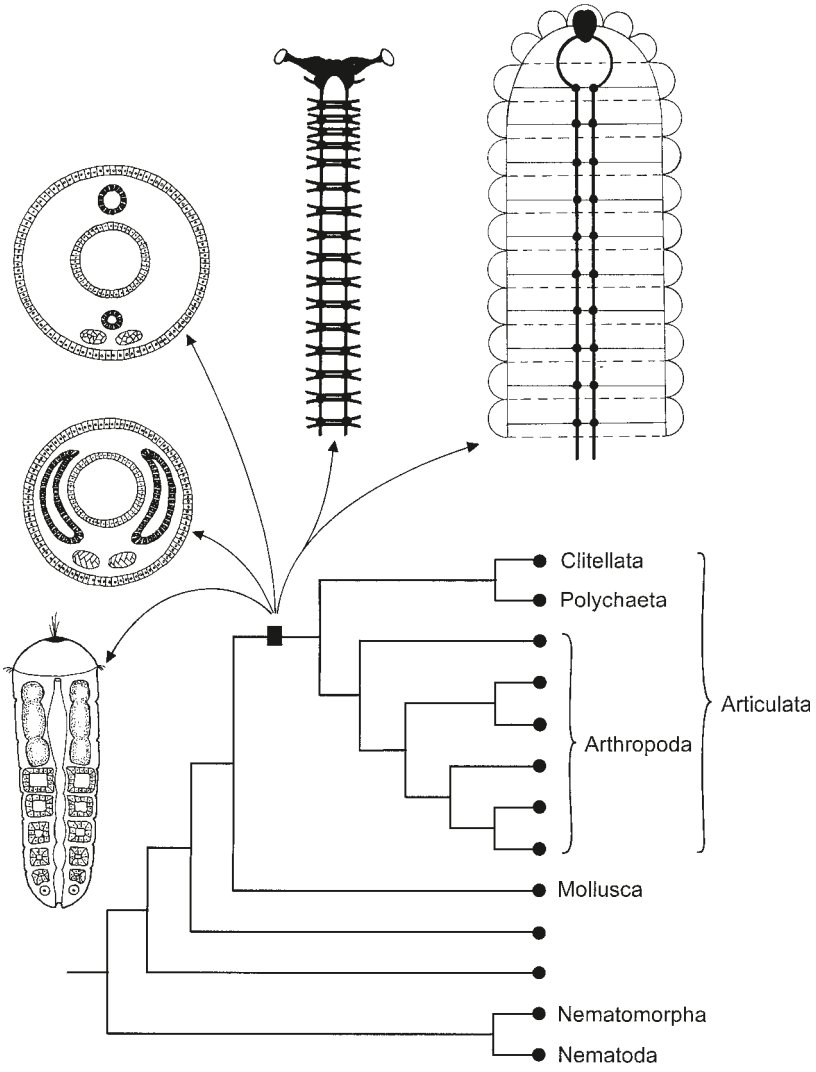


Fig. 174. The Articulata hypothesis is supported by several morphological and ontogenetic similarities shared by arthropods and annelids (discussed e.g., in Ax 1999, Westheide & Rieder 1996, Brusca & Brusca 2003), which are complex enough to be considered as important apomorphic homologies. These include the ontogenetic preanal formation of segments, the segmental coelomic sacks (also in arthropod embryos, they disappear at a later phase), the dorsal position of the tube-like blood-pumping vessel ("dorsal hearts") and the direction of blood-flow (anteriorly), the structure of the central nervous system (the ladder-type nerve system of a species of Annelida and a species of Arthropoda is depicted here; the ground pattern of the nervous system of the Articulata still has to be reconstructed in detail), the segmental anlage of the nephridial organs (not shown). All characters listed are lacking in Nemathelminthes (only larger taxonomic groups also included in Fig. 172 are shown).

ing datasets contains better characters. Such a comparison is difficult because usually there exists no information on the genetic basis for morphological character state changes that could be quantified in the same way as substitutions. In the literature there exist several examples for untrustworthy hypotheses founded on molecular data that support topologies incompatible with usually older but nevertheless convincing information derived from comparative morphology (Fig. 172, 174), and there exist also insufficiently justified traditional hypotheses which were con-

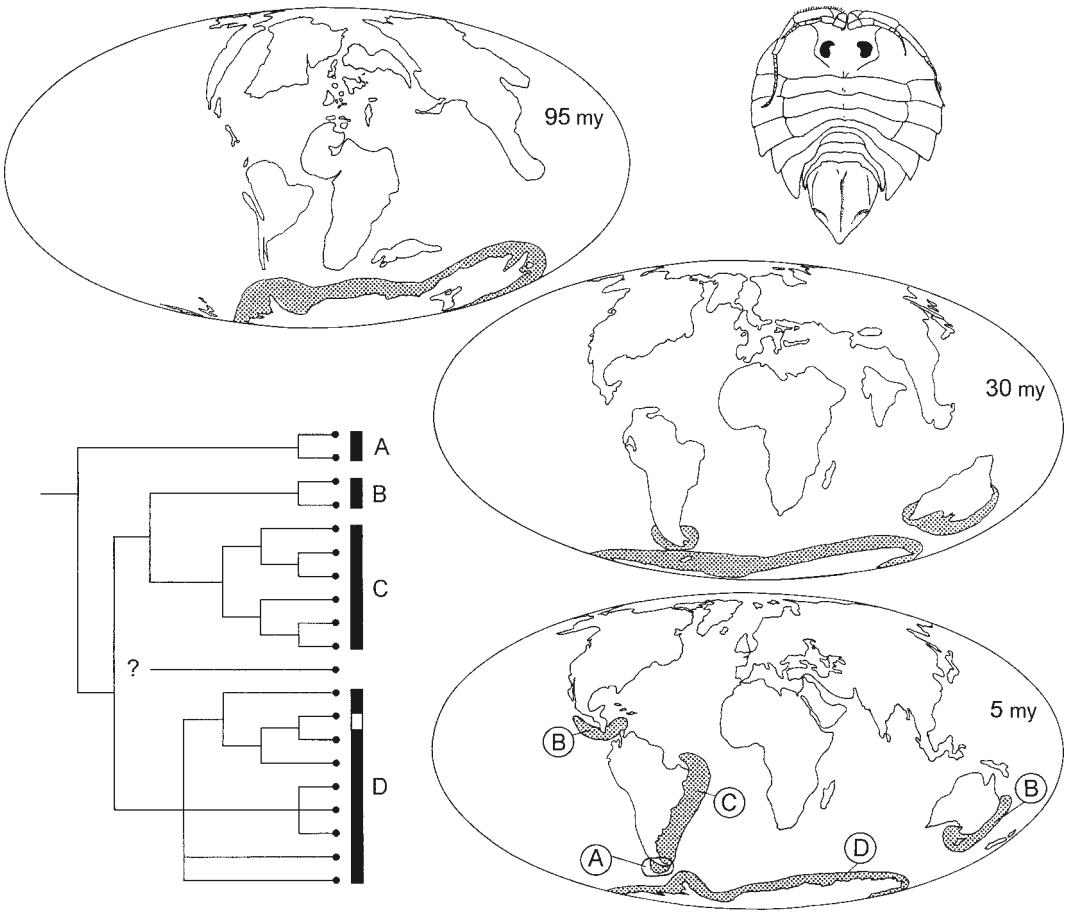


Fig. 175. Congruence between distribution patterns and phylogeny of the Serolidae (Crustacea Isopoda, after Wägele 1994). After the separation of Africa from the original Gondwanan continent, the ancestors of modern Serolidae must have evolved in the colder waters along the shores of southern Gondwana. The more archaic serolids (group A) still live in Patagonia today. Australia separated next from Antarctica, South America followed about 35 million years ago. The groups of genera B and C show local radiations in Australia and South America. After the separation of South America, the Southern Ocean around Antarctica cooled down. Populations adapted to the polar climate (group D) and a further local radiation took place around Antarctica. It is not clear why species of Central America are similar to Australian ones. Hatched regions in the two upper maps show the probable distribution of the last stem lineage representative of recent serolids.

tradicted by new molecular data (examples: monophyly of Nemathelminthes s.l. (including e.g., Rotifera) and of Crustacea were never based on convincing apomorphies).

The superficial similarity of body and head form of new world vultures and old world vultures had the consequence that the condor and its South American relatives (Cathartidae, Fig. 70) were for a long time classified as birds of prey (Falconiformes) (e.g., Mauersberger 1974). However, anatomical as well as independent molecular data showed a closer similarity to storks (Ciconiidae: e.g., König 1982, Wink 1995, Sibley & Ahlquist 1990). The contradiction between the classification based on the superficial similarity in shape and life form and the results of phylogenetic analysis can be explained by adaptations to scavenging which originated convergently in South American birds and old world vultures.

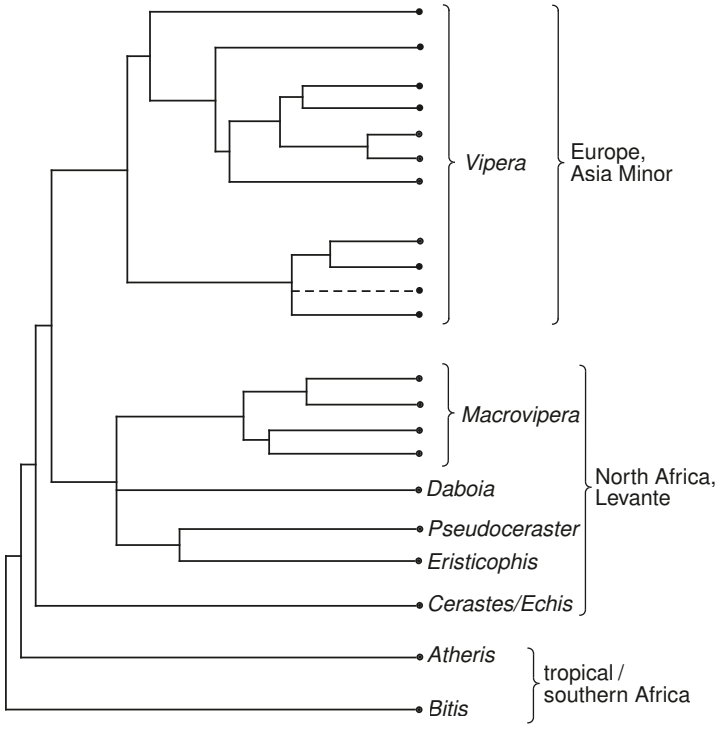


Fig. 176. Of two hypotheses on the phylogeny of vipers one assumes the polyphyly of palaearctic and North African species groups (Ashe & Marx 1990), the other (depicted) assumes monophyly (Joger 1996). The latter is more plausible considering the geographic distribution.

Congruence between different datasets increases confidence in a topology. But it is also possible that the same wrong phylogenetic tree is obtained several times independently. One cause for this can be the rapid evolution and radiation of a group of species, which causes a substitution of apomorphic states that once existed in the last common ancestor (erosion of synapomorphies). It may happen that with repeated phylogenetic analyses the monophyly of this group is not detectable. More conserved plesiomorphic characters would support paraphyletic groups. A conspicuous example is the sistergroup relationship between Annelida and Mollusca under exclusion of Arthropoda which for this reason has been found several times in molecular analyses, while a large number of anatomical features indicate that arthropods had an annelid-like ancestor. As in the recent fauna there is a large gap between annelids and arthropods, and because arthropods are very diverse morphologically and also derived at the level of genes, it is not possible to reduce for a molecular analysis the large genetic distance between these groups by addition of further species to shorten the long branch leading to arthropods.

Historical biogeographic patterns

It can be expected that whenever the mobility of organisms is small, more closely related species also occur in spatial proximity. In case new species evolved while populations dispersed it should be possible to project phylogenetic trees on distribution patterns (Fig. 175, 176, 177). Deviations require additional assumptions: when a species of a monophylum occurs at great distance from the other members, either exceptional events increased the mobility of some specimens (storms, transport on drift wood, anthropochory; e.g., Galápagos finches), the populations might have been separated by continental drift (relatives of ostriches in South America, Africa and Australia), or the species is a survivor in an originally larger area of distribution (for example, tapirs in

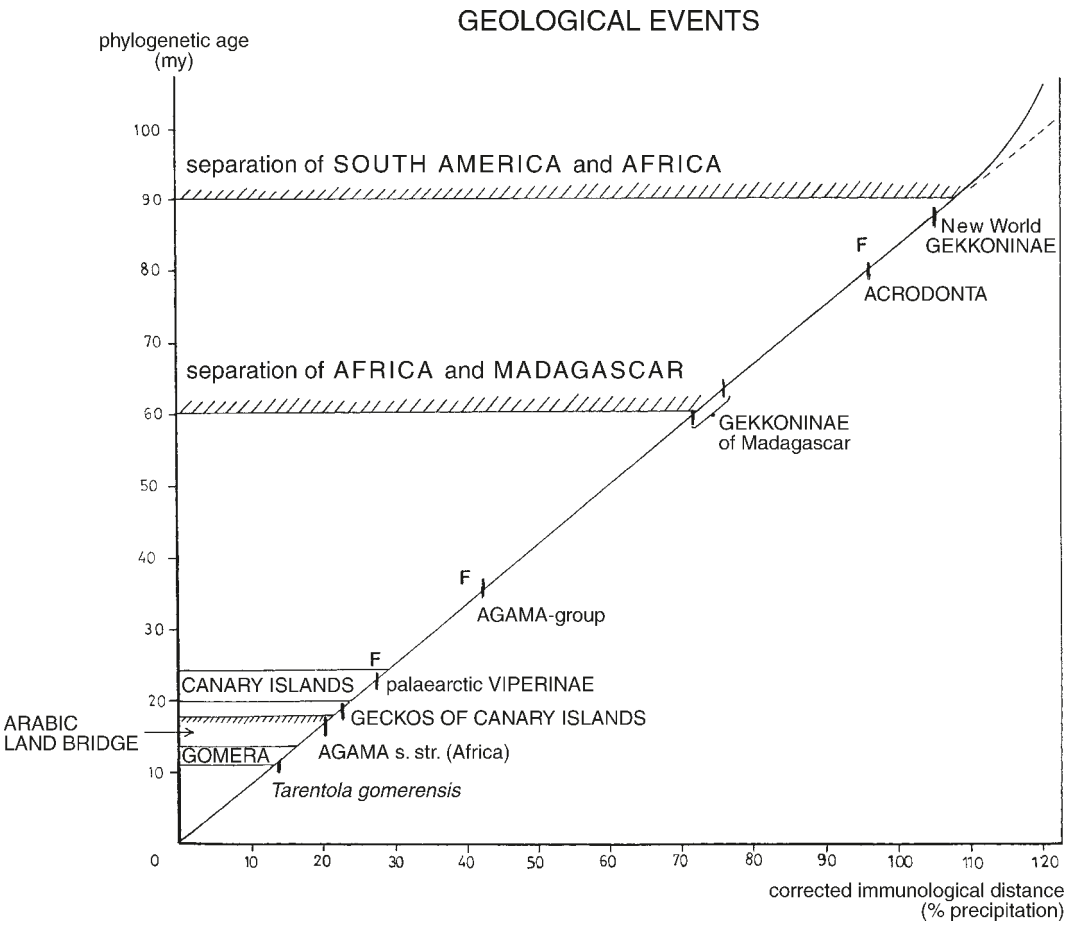


Fig. 177. Immunological distances of spatially isolated taxa of reptiles to the respective sister taxa. Differences of immunological distances of albumins are congruent with the occurrence of fossils and age differences of geological events causing vicariance (after Joger 1996).

South America and in Malaysia). Motive to question hypotheses of phylogeny and dispersal arises if the explanation for the cause of the actual distribution appears to be improbable, for example, if ocean currents are predominantly opposite to the postulated dispersal route of marine species.

The ecological development of landscapes is always linked with climatic and geological history. Orogenesis and oscillations of sea level have consequences for climate, vegetation and animal communities. A phylogenetic analysis is plausible if inferred speciation events can be brought into accordance with such changes of climate and the biotic environment.

In Fig. 178 the assumed separation and divergence of populations is in accordance with historical processes causing climate changes and the development of distribution barriers (Cracraft 1983). In the area of the Gulf of Carpentaria (Northern Australia), an eastern and a western region were separated by a decrease of precipitation (barrier A, the area in the east is more humid). Afterwards in the West a further separation of regions occurred through the development of river beds (barrier E) and dry areas. Barriers B and C also separate climatic zones and vegetation types.

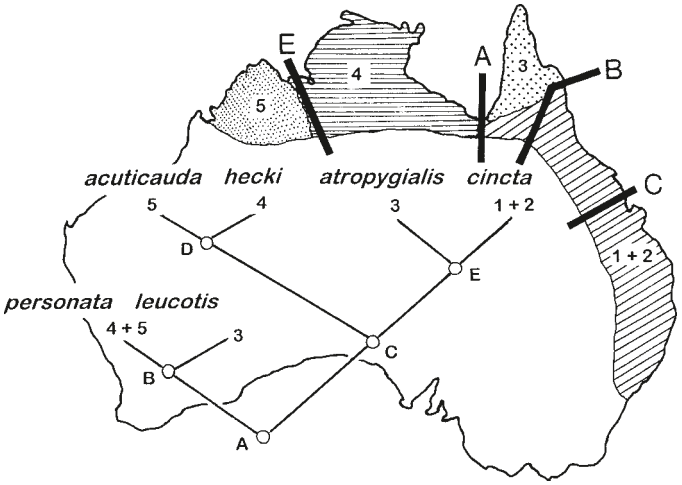


Fig. 178. Areas of Australian bird species (genus *Poephila*) and their assumed phylogeny. The bars represent distribution barriers, the numbers on the dendrogram refer to distribution areas (see text). Barrier C is a barrier for other bird taxa (letters on nodes of the tree do not refer to barriers; after Cracraft 1983).

Fossils

Newly discovered or at first disregarded fossils can contribute to verify phylogenetic hypotheses. They provide evidence on

- former areas of geographic distribution,
- the chronological succession of character states,
- the minimum age of taxa.

Examples: The marine Serolidae is a taxon of Isopoda (Crustacea) distributed in the Southern hemisphere and characterized by a disc-shaped body outline. Contrary to the traditional view, a phylogenetic analysis of the Isopoda suggested that the disc-shaped body is not an autapomorphy of Serolidae but a character of the ground pattern of the higher ranking taxon Sphaeromatidea (Wägele 1989). The latter also includes the Sphaeromatidae, most of which can roll up when disturbed. The fossil *Schweglerella strobli* Polz, 1998 (which is not a member of Serolidae) proves that the serolid-like disc shape was also present in other less derived species of Sphaeromatidea that lived outside the Southern hemisphere: *Schweglerella strobli* was discovered in Solnhofen (Southern Germany) (Fig. 179). – Marsupials are a characteristic element of the Australian fauna, today including well known animals like kangaroos, koalas, and wombats. The Didelphoidea, which are native to South America, are morphologically more primitive and are considered to be phylogenetically older than the Australian taxa (e.g., Carroll 1993). Plate tectonic events and fossils are in agreement with this theory: in the upper Cretaceous South America, Australia and Antarctica were not separated and there existed at least a chain of islands between North and South America. The oldest fossils are known from North America, from where migrations to South America and Australia were possible via Antarctica. A fossil discovered in Antarctica (Woodburne & Zinsmeister 1984) proves that the Antarctic continent was also once colonized and supports the hypothesis of the descent of the Australian marsupials from opossum-like American ancestors.

Example for the reconstruction of character evolution: insects, myriapods and crabs are often combined in the taxon "Mandibulata" because they have the same head composition with 3 pairs of mouthparts (1 pair of mandibles, 2 pairs of maxillae). It is assumed that the last common ancestor already had mandibles and two maxillae (e.g., Snodgrass 1950, Wägele 1993, Scholtz et al. 1998). New Cambrian fossils show that primitive stem lineage representatives of the Mandibulata at first also included the second antenna in the mouthparts (Fig. 99). In more derived animals the following appendages became also specialized, however at first only the mandible and the first maxilla. Since also within the Crustacea the second maxilla of the Cephalocarida is not

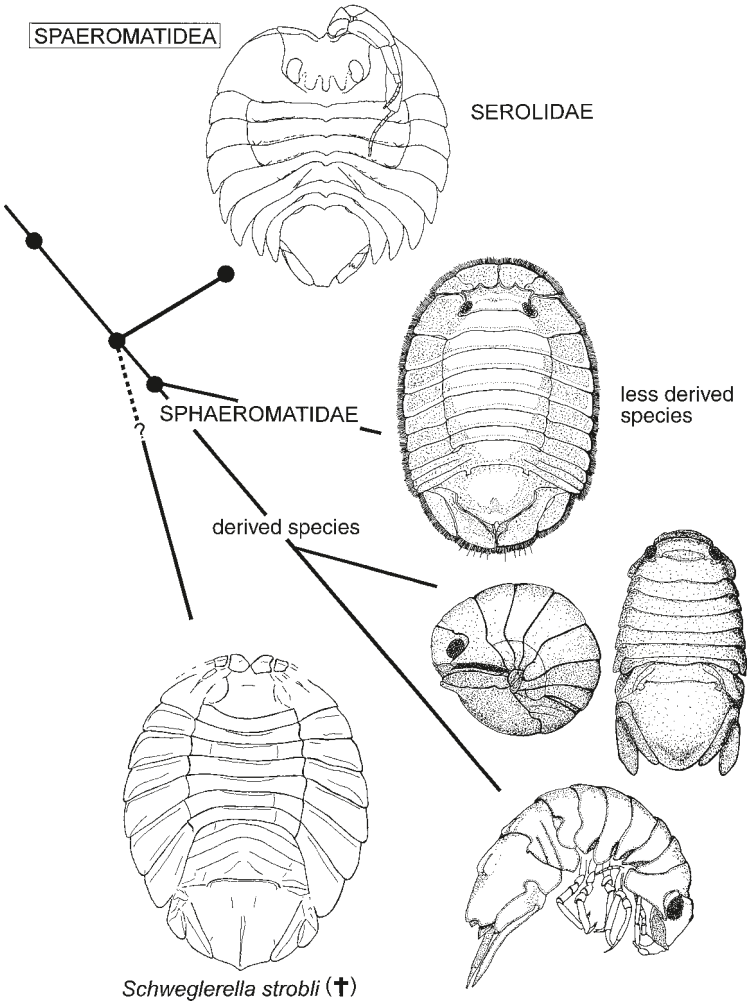


Fig. 179. Verification of a phylogenetic hypothesis with fossils. A comparison of the morphology of different marine relatives of wood lice (Isopoda) led to the hypothesis that Sphaeromatidae with a disc-shaped body are the less derived forms within the family and that this body shape is homologous to that seen in Serolidae (Wägele, 1989). The discovery of the fossil *Schweglerella strobli* from limestone of Solnhofen (Polz 1998), an animal that belongs neither to the Sphaeromatidae nor to the Serolidae, proves that this body form is not unique to these families and is a character of the ground pattern of a higher ranking taxon (Sphaeromatidea; not all taxa of the Sphaeromatidae are shown here). The plesiomorphic state within the Sphaeromatidae is confirmed with this discovery.

specialized and looks like a thoracic leg, the concept of a ground pattern of the Mandibulata with 3 pairs of specialized mouth parts has to be revised (the alternative is to assume an atavism in Cephalocarida). In this case, the second maxilla must have evolved to a mouthpart within the Mandibulata. This hypothesis can be verified if this scenario of character evolution can be mapped on the preferred phylogeny of the Mandibulata.

Character evolution and ways of life

An analysis of the 18S rRNA genes of marine isopods (related to wood lice) showed that the Bopyridae, which parasitize crabs, are closely related to the Cymothoidae, animals that suck blood on fish, and verifies earlier morphological analyses (see Wägele 1989). These results of phylogenetic studies are plausible because they are

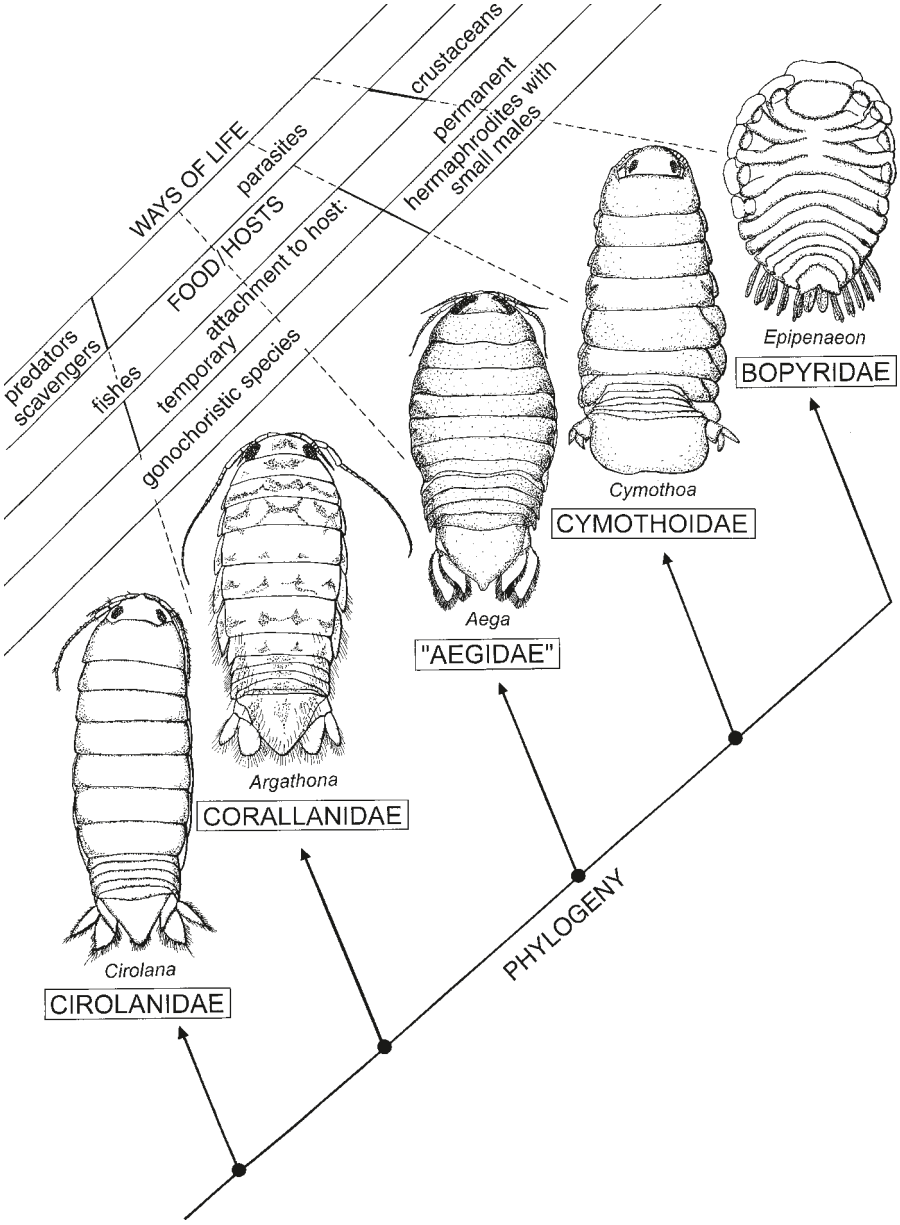


Fig. 180. The gradual specialization scenario for the evolution of parasitic isopods inferred from modes of life of recent species (evolution of modes of feeding, host preferences, life cycles and hermaphroditism) is in agreement with the results of phylogenetic cladistic studies of morphological characters and molecular analyses of 18S rDNA sequences (Dreyer & Wägele 2001).

in agreement with the most probable scenario for the evolution of modes of life (Fig. 180): in both taxa the mouthparts are adapted to piercing of host tissues, and the same appendage parts are modified in comparison with outgroup taxa. In both parasitic taxa additionally the pereopods 1-7 have grasping claws used to cling to the host. Furthermore, the life cycles are similar: juveniles swim fast searching for hosts, adult animals are sessile and morphologically highly specialized.

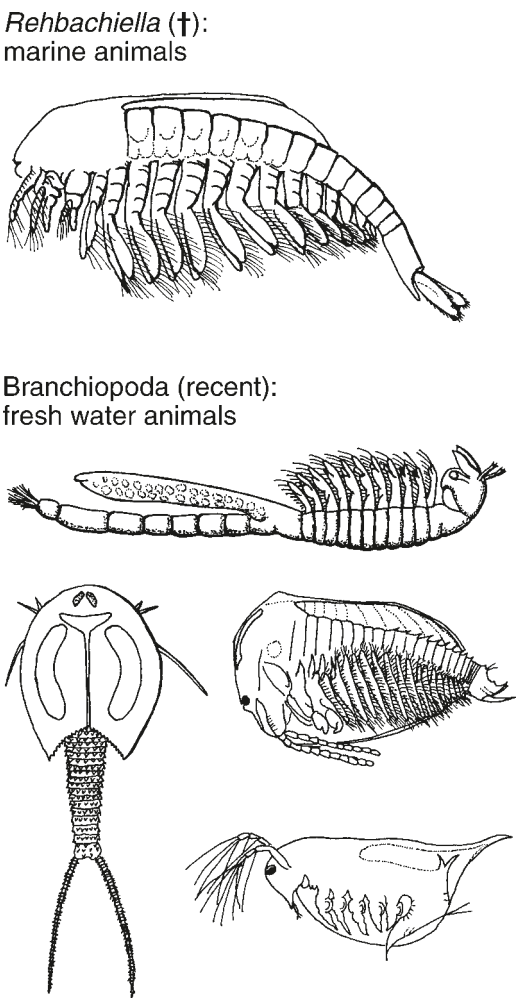


Fig. 181. Branchiopoda are crustaceans which are adapted to ephemeral epicontinental waters. The Cambrian marine fossil *Rehbachiella* raises new questions on the homology of these adaptations (see text).

The phylogeny implied by the dendrogram obtained for these taxa allows the statement that evolution of parasites started with carrion feeders which prefer fish (Cirolanidae), which gave rise to ectoparasites that suck blood on fish (e.g., Aegidae), and a further evolutionary specialization led to parasites living permanently on fish in the adult stage (Cymothoidae). The host change from fish (hosts of the Cymothoidae) to crabs (hosts of the Bopyridae) can also be explained, as several adult Cymothoidae occasionally suck hemolymph of crustaceans.

The explanation for the evolution of modes of life is not a type of unfounded "story telling" but the most parsimonious connection of different modes of life (requiring the lowest number of new adaptations from clade to clade). Modes of life are based on genetic properties, but since these are usually not known (what is causing dwarf males and protandric hermaphroditism?) these characters are not used for cladistic analyses.

Non-trivial hereditary adaptations to environmental conditions are characters which often are not identified for phylogenetic analyses, but they are nevertheless present. These include for example adaptations to fresh water in species of marine origin, or adaptations to terrestrial life, to dry habitats, polar climates, specific types of food. These adaptations can be assessed as putative apomorphies, and they can be employed for tests of plausibility. If a hypothesis of relationships implies an evolution of adaptations and ways of life that is not parsimonious or that seems to be unlikely (not as in Fig. 180), this hypothesis has to be re-examined in the light of new characters or new character analyses, or alternative explanations for the evolution of modes of life are required.

Example (Fig. 181): amongst crustaceans the Branchiopoda (consisting of Anostraca and Phyllopoda) are specialized to colonize ephemeral epicontinental waters. They have hard-shelled eggs which can desiccate and rest dormant in the soil for years. When water is added to a suitable sample, after a few days the larvae (nauplius stage) hatch which develop rapidly. The animals possess variations of a filtering apparatus formed by special appendages (phyllopods) suitable to collect plankton. As archaic and defenceless animals, they survive especially in ephemeral habitats which are not accessible to other competitors and many predators like fish. It can be inferred from morphology and the mode of life that the group is monophyletic, and it is highly probable that this mode of life was already part of the ground pattern. This concept is in conflict with the classification of the Orsten fossil *Rehbachiella*, which was found in marine sediments and was considered to be a Cambrian member of the Branchiopoda (Walossek 1993). Leaving aside problems with the evaluation of morphological characters, the marine habitat would not require

dormancy of draught-resistant eggs. With the discovery and interpretation of this fossil new hypotheses arise which have to be tested: (a) *Rehbachiella* is not a branchiopod, (b) *Rehbachiella* is a secondary marine species derived from an epicontinental fresh water ancestor, or (c) the adaptations described above are convergences within the recent branchiopods. These hypotheses can be partly tested with more detailed character analyses.

10.2 Comparison of topologies

Attention: the following tests do not estimate the quality of data or the plausibility of results, and they are not suitable to check if a model mirrors the true evolutionary processes. They allow the comparison of trees or datasets within the framework of ML analyses. The three tests describe if differences in topology are significant when a single dataset is given. The tests are designed to compare two strictly *bifurcating* trees.

Kishino-Hasegawa test (KH-test)

If two different trees are to be compared and some data are available that are assumed to be a good sample from the real phylogeny, it is interesting to know if the difference in topology (not in branch lengths) is statistically significant. The likelihood ratio test (ch. 8.1) can not be used, because the difference in degrees of freedom between two trees is not known and the hypotheses are not necessarily nested (required for chi-square distribution of the likelihood ratio test statistics). With the likelihood-based method developed by Kishino and Hasegawa (1989) one can estimate standard error and confidence intervals, however, it is required that the trees are specified *a priori*, that a correct substitution model is known, and that the data are a representative and independent sample from the true phylogeny. The null hypothesis is that the two tested trees are not different.

Assuming that two topologies T_1 and T_2 have been selected *a priori*, one can estimate their likelihoods L_1 and L_2 (see ch. 8.3 and 14.6) and calculate the difference $\delta = L_1 - L_2$. Goldman et al. (2000) suggest to proceed in the following way: resample the data (non-parametric bootstrapping), re-estimate the likelihoods for T_1 and T_2 for each replicate i optimizing free model parameters and then calculate $\delta^{(i)} = L_1^{(i)} - L_2^{(i)}$. The difference between $\delta^{(i)}$ and the mean of $\overline{\delta}^{(i)}$ of all replicates is $\widetilde{\delta}^{(i)}$ (centering procedure). The resulting set of values $\widetilde{\delta}^{(i)}$ gives an estimate of the distribution of δ under the null hypothesis. Rank these $\widetilde{\delta}^{(i)}$ values, define a confidence interval (for example between 2.5 % and 97.5 % of the ranked list) and check if the value of δ calculated from the real data falls within the confidence interval.

This procedure is time-consuming, because for each bootstrap replicate all model parameters are estimated from the data. An approximative method that performs well is to use the model parameters obtained from the original dataset. Other variants are discussed by Goldman et al. (2000).

The null hypothesis of this test (the trees are not different) is only justified if the topologies were selected without reference to the data that are used for the test. An incorrect usage of the KH-test is to compare a tree that has a maximal likelihood for a given dataset with another tree, or to compare the optimal ML-topology with suboptimal ones using the same dataset. In the latter cases one naturally cannot expect that the difference between the trees is insignificant. Unfortunately, the KH-test has often been used in exactly these unsuitable cases. To compare the likelihood scores of trees derived from some data at hand the Shimodaira-Hasegawa test is more appropriate.

Shimodaira-Hasegawa test (SH-test)

This test resembles the KH-test and also needs a preselected substitution model, but it allows multiple comparisons of trees (Shimodaira & Hasegawa 1999). It requires a set M of topologies that contains every topology that might be entertained as the true topology and a condition is that the selection of topologies for the set M is made

a priori without reference to the data used for the test. The null hypothesis is that all trees contained in M are equally good explanations of the data. Goldman et al. (2000) propose the following procedure:

Calculate the maximum likelihood topology T_{ML} for the dataset. Calculate the likelihood difference δ_x for the trees T_x in the set M ($\delta_x = L_{ML} - L_x$) with the dataset that is being used for the test. Bootstrap this dataset and maximize the log-likelihood $L_x^{(i)}$ for each bootstrap replicate i and for each topology T_x. The mean likelihood $\bar{L}_x^{(i)}$ of all trees obtained with replicate i is subtracted from each value $L_x^{(i)}$ to get the adjusted value $\tilde{L}_x^{(i)} = L_x^{(i)} - \bar{L}_x^{(i)}$ (centering method). $\tilde{L}_{ML}^{(i)}$ is for each replicate i the value $\tilde{L}_x^{(i)}$ of the tree T_x that has the maximum adjusted likelihood difference. For each bootstrap replicate i and for each topology T_x calculate the difference $\delta_x^{(i)}$ with $\delta_x^{(i)} = \tilde{L}_{ML}^{(i)} - \tilde{L}_x^{(i)}$.

Define a confidence interval for $\delta_x^{(i)}$, for example between 0 and 95 % of a list of ranked $\delta_x^{(i)}$ values (one-sided test, significance level 5 %). If for a topology T_x the attained δ_x falls within the interval, it can be considered to be a sample from the distribution. The SH-test compares the topologies T_x and T_{ML} on the basis of differences of the topologies in explaining the observed data.

SOWH-test

This test (Swofford et al. 1996, Hillis et al. 1996) relies on **parametric bootstrapping** (ch. 6.1.9.2) and it allows to find out if a topology T_1 that had been selected *a priori* is supported by a dataset. The following description is based on Goldman et al. (2000):

Calculate the difference in likelihood between the selected tree T_1 and the optimal tree T_{ML} with $\delta = L_{ML} - L_1$. Simulate datasets evolving artificial sequences of random composition along the topology T_1 and with model parameters derived from the original data for T_1. Re-estimate with these replicate datasets i the free model parameters using T_1 and calculate with these parameters the log-likelihoods $L_1^{(i)}$. Estimate for each replicate the optimal ML-tree (which usually differs from T_1) and get for it the likelihood value $L_{ML}^{(i)}$. The difference in likelihoods between T_1 and T_{ML} is $\delta^{(i)} = L_{ML}^{(i)} - L_1^{(i)}$.

Define a confidence interval, for example between 0 and 95 % of a list of ranked $\delta^{(i)}$ values (one-sided test, significance level 5 %) and see if δ falls in this interval.

11. The importance of results of phylogenetics for other studies

As a result of a phylogenetic analysis four classes of hypotheses are obtained:

a) Hypotheses of phylogeny (data on sistergroup relationships and on the composition of monophyletic groups).
b) Hypotheses on the direction of character evolution and identification of the stem lineage in which a novelty evolved: these are only inferable for discrete characters. With increasing divergence time more complex characters are better suited for the determination of character state polarity, because states can be homologized with better confidence. Distance trees do not analyse character transformations directly. However, comparing a distance topology with a character matrix, the distribution of each character state can be mapped on the tree and node characters can be inferred (e.g., as in Fig. 111).
c) Hypotheses on the speed of character evolution: these can only be gained with methods that allow the evaluation of quantitative characters (ideally the number of substitutions generating a character), and in addition when divergence times can be estimated (ch. 2.7.2.3). Morphological characters are not suitable for this purpose when their genetic background is unknown.
d) Hypotheses on the correlated evolution of different characters.
e) Hypotheses on the ground patterns of taxa. These yield information on genes and functions that probably were once present in ancestors.

These hypotheses can be the indispensable basis for further studies:

- analyses of the adaptation of organisms to their environment, distinction between convergent adaptations triggered by environmental parameters and homologous adaptations,
- analysis of the inheritance of properties,
- analysis of the dispersal paths of species (historical biogeography),
- analysis of the historical factors which imprinted the biosphere and analysis of the age of ecosystems,
- estimation of the period of time in which species diversity evolves,
- analysis of the factors causing and maintaining species diversity,
- prediction of the properties of species that are not well known, based on knowledge about their relationships and with data on properties of related species (e.g., search for genes of commercial interest, products relevant for industry, medicine or agriculture),
- distinction of species and races (e.g., pathogens, contagious agents, vectors, plant pests, new varieties for agriculture, climatically adapted beneficial organisms for local projects, analysis of food chains in ecosystems, protection of species diversity),
- furthermore the systematization of species is essential to order, classify, store, and recall biological knowledge.

Furthermore, phylogenetic studies can have a great economic interest (identification of microorganisms relevant for medicine, agriculture, biotechnology, search for relatives of known species relevant for fisheries, agriculture or the pharmaceutical industry, evaluation of the biodiversity in interesting plots of land or areas of the ocean).

12. Systematization and classification

The classification of objects that is essential for our colloquial language (classification of plants according to size, cars according to their use) generally does not require a scientific theory. Remember that usually a classification is the grouping of objects according to properties that are chosen subjectively. A classification is performed with predicators, not with proper names (ch. 1.2).

As long as the classification of organisms does not correspond to their phylogeny one can talk of a *phenetic classification*, which means a grouping according to similarities. In biology, however, a classification of organisms that mirrors aspects of the topology of the phylogenetic tree is required. The filing of organisms into the phylogenetic system and the naming of monophyla is called *systematization* (s. ch. 1.3.4). This is a classification with the help of a theory-dependent system. A systematization is not exclusively carried out according to properties of objects, but on the basis of the reconstructed descent. Descent is *not a property* of objects, but a historical process which can be reconstructed with the help of identified homologies. Therefore, systematization is a special variant of classification.

12.1 Systematization

A result of a phylogenetic analysis is the distinction of **monophyletic groups** which are arranged encaptically or as sistergroups. For the description or graphical representation of the system it is helpful to use proper names for the monophyla and statements or graphs explaining their genealogical relationships.

A genealogical order can be explained with dendrograms, Venn-diagrams (ch. 3.2), or in words. To name all monophyla which are visible in a dendrogram is neither necessary nor desirable. A classification represented by proper names has to be rejected when a Venn diagram of monophyla is not compatible with the Venn diagram representing the hierarchical order of proper names used for the species considered (Fig. 182).

One may ask why a biological system should be composed of monophyla. Alternative concepts could allow the inclusion of polyphyletic or paraphyletic groups. What are the disadvantages? To answer this question it is necessary to define the desired properties of proper names of taxa:

- the relation between a name and the mental grouping of organisms represented by this name should be unequivocal,
- this relation should be stable,
- the system should reflect phylogeny,
- the system should have a heuristic value.

Concerning the definition: monophyla can easily be defined unequivocally with different methods (see ch. 4.4.1). To delimit a paraphylum one has to name the last common ancestor and in addition all descendant taxa that are excluded from the group. To delimit a polyphylum one has to name more than one common ancestor of part of the group members and in addition all descendant taxa that are excluded. It is therefore easier to work with monophyla.

Concerning stability: The number of groups that can be defined from a given phylogeny increases in the order monophyla < paraphyla < polyphyla. Furthermore, using only monophyla it is not allowed to pick out of a group a species that shows some derived feature and to create for this species a new taxon of the same rank as the "source taxon". Using paraphyla or polyphyla this is allowed and everybody can create new taxa and contribute to an increasingly confusing system. Therefore, such systems are less stable and complicate communication.

Reflection of phylogeny: if a phylogeny is given, different scientists can propose different classifi-

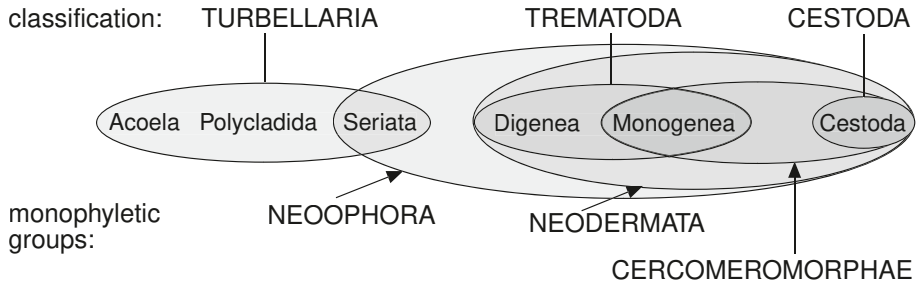


Fig. 182. Incompatibility between a classification and a phylogeny in the case of the Plathelminthes (classification according to Remane et al. 1996, simplified systematization according to Ehlers 1985). The phenetic classification follows morphological similarities and ways of life: the Turbellaria are free living and ciliated species, the Trematoda are ecto- or endoparasitic as adults and have a neodermis without cilia, suckers, and a normal gut, whereas the adult Cestoda are non-ciliated, ribbon-like animals without a gut.

cations. If only monophyla are allowed, all proposed taxa will always fit to a perfect encaptical (hierarchical) order, without overlap in Venn diagrams representing this order. If paraphyla and polyphyla are used, overlap of taxa in alternative classifications is inevitable. – If monophyla are used, the whole system can be constructed with little additional information ("Who is the sister-group of each taxon?"). Otherwise, even if no overlapping taxa exist, additional information is required ("Which taxa are excluded from a group and at which branches within a group's local tree must the next group be connected?"). Furthermore, to avoid misunderstandings, it is necessary to mark clearly monophyla, paraphyla, and polyphyla in a different way. Therefore, a system based on monophyla is less susceptible to errors and misunderstandings.

Heuristic value: if only monophyla are used, one can predict that a new species of this monophylum will show many groundpattern characters and especially the apomorphies of this monophylum with high probability. This probability is the same in paraphyla but much lower in polyphyla. Furthermore, predictions about the placement of a new species within the system are only reliable when the system is composed of monophyla: the argument "this is with high probability a mollusc because the animal has a radula" is less reliable when molluscs are para- or polyphyletic.

12.2 Hierarchy

The hierarchy of proper names or of taxa within a group of organisms is nothing else but a representation of the encaptic order of monophyletic groups. As in taxonomy there exist no hierarchical levels which refer to real units of time, to the age of taxa, their extent in time, or to some processes that are relevant for all organisms, it is not possible to assign ranks objectively. Single ranks have no relation to phenomena of nature ("the genus" does not exist in nature).

Therefore we must state that a category "family" or "suborder" is not comparable to the military ranks "general" and "colonel". The latter have comparable qualities, at least comparable rights and duties, independently of the single real individuals that are awarded the rank. In biology, however, a "family" of flies comprises many more species and other divergence times than a "family" of mammals (see also Fig. 66), a comparable quality does not exist.

12.3 Formal classification

Rules for the classification and naming of organisms are necessary to achieve stability: nobody wants to learn again and again new names for groups of animals and plants or for the same species. It is desirable to keep the relation between a name and a group of organisms as constant as possible. However, in practice some flexibility is required because scientists can make mistakes either because they miss an earlier publication and propose a name for a clade or species that is already named or because they did not uncover the correct phylogenetic relationships.

12.3.1 Traditional Linnéan nomenclature

The formal classification is determined by the rules international commissions (ICNB 1992, ICBN 1994, ICZN 1999). With these rules the formal naming of taxa, the priority of homonymous and synonymous names and the use of some Linnéan categories is stipulated. Until recently, international rules implied that taxon names are always coupled with categories although objective criteria for the assignment of ranks do not exist. Endings for taxon names depend of the rank and are dictated in zoology up to the level of families or superfamilies. Example:

Class Arachnida
Order Araneae
Suborder Opisthothelae
Superfamily Araneoida
Family Tetragnathidae
Genus *Tetragnatha*

The rules that are currently valid dictate some widely accepted formalisms:

- the binominal species name (composed of the genus name and an epithet, as in *Homo sapiens*),
- rules for the formation of scientific names (e.g., the family name referring to the genus *Coccinella* is obtained using the stem Cocinell- and adding the suffix -idae, resulting in Coccinellidae),
- for synonyms and homonyms the precedence of older valid names (*synonyms*: two or more names denoting the same taxon, *homonyms*: names with the same spelling denoting different nominal taxa, a *preoccupied name* is a *junior homonym*),
- the type concept (type specimens are the reference for species names, type species for genera, type genera for families),
- the hierarchical order is described with Linnéan categories,
- rules for the validity of a publication.
- Names can only be valid if they were published after 1753 (plants) or 1758 (animals).
- Descriptions must be written in the Latin alphabet.

One important principle is that of *priority*. If the same species or higher ranking taxon has been described several times with different names, the first name applied to a taxon is the valid one. This prevents debates about the correct name of a taxon.

In their present state these rules do not consider the laws of phylogenetics! The traditional nomenclature does not indicate

- how to select a category,
- which species or subordinate taxa are to be included in a new taxon,
- whether taxa should be monophyletic or not,
- how to adapt the content of names if hypotheses on relationships change,
- how to write an unambiguous diagnosis for a taxon,
- if splitting or lumping of taxa is recommendable or reprehensible.

The result is that different names are used for the same clade (e.g., Bopyrinae and Bopyridae) depending on the assigned category, or the same name is used for different clades (for example, after splitting of a large clade into smaller units). Therefore, a systematist should voluntarily keep to additional rules (see below) to contribute to a greater stability. For the time being it is often not possible to publish descriptions of newly discovered species or species groups without suggesting a formal classification with the assignment of Linnéan categories due to the established norms observed by editors of scientific journals. However, taxonomists should primarily take the following more important rules into account:

- Each taxon has to be monophyletic, because this is the only way to delimit taxa objectively and based on empirical knowledge.
- If the assignment of a species to a taxon is uncertain because available information is insufficient, the next higher ranking taxon that shares synapomorphies with the problematic species should be chosen wherein the species can be placed as *incertae sedis*. A mud-shrimp whose placement in known families could not been clarified can be placed into the corresponding superfamily or into the suborder (Thalassinidea *incertae sedis)* because it shows apomorphies of the Thalassinidea.
- In formal descriptions of a new taxon the apomorphies of the taxon should be stressed and distinguished from diagnostic characters and plesiomorphies. The discussion of these apomorphies must be contrasted with the plesiomorphic character state seen in outgroup taxa.
- When in the course of revisions larger taxa are split or new groupings are proposed, the erection of new taxa should be based on phylogenetic analyses. The proposed phylogeny should be illustrated graphically.
- After identification of subtrees within a named monophylum, the name of the monophylum should not be used to name a subtree (an unnecessary splitting of a named taxon that produces a homonym).
- The splitting of well established monophyletic taxa into smaller units with the aim to create new names should be avoided whenever the creation of new names is not coupled with a gain of knowledge or if it does not improve communication or handling of taxa.
- When a subgroup is separated from a larger monophyletic group of species and named, the rest of the larger group should not remain as a paraphyletic taxon. If the phylogeny of the paraphyletic group is not resolved, the members have to be kept as *incertae sedis*. The better solution is to resolve the complete phylogeny and to name only monophyletic groups.
- When the assignment of categories becomes necessary under pressure of journal editors, the hierarchy should only follow the tradition; the order of ranks should reflect the encaptic order of the phylogenetic system.
- The naming of supraspecific monophyla is only convenient and helpful when groups are delimited which can be distinguished easily by taxonomists or that have apomorphic features important for understanding evolutionary, ecological or physiological processes. An inflation of names does not serve anybody.
- When the order in which taxa are listed is chosen to represent the chronological series of speciation events, one should mention this intention explicitly because most taxonomic lists have no relation to phylogeny. The sequencing convention implies that names listed at the same level of indentation are sistergroups, more indented names are subgroups.
- Categories have often been used for long known groups of recent species, which is why for the hierarchical level of newly discovered monophyla no Linnéan categories are left. Clades composed of fossil species which have to be placed on a stem lineage and that together with recent species form a monophylum *M* for which no Linnéan categories are available (e.g., between the categories superfamily and family) can be classified with the variable and rankless category "**plesion**", avoiding invention of an additional category, and also the larger monophylum *M* can get the category plesion (Patterson & Rosen, 1977). The corresponding sistergroup relationships can be visualized with a dendrogram or with a written sequential list. The following list corresponds to the phylogeny of mammals, which starts with mammal-like reptiles (after Carroll 1993: 376); the first taxon comprises the following ones:

  Plesion Pelycosauria
  Plesion Therapsida
  Plesion Cynodontia
  Class Mammalia

In this list Therapsida are a taxon of Pelycosauria, Cynodontia a taxon of Therapsida, Mammalia a taxon of Cynodontia.

The use of the category plesion is not recommended because it requires arbitrary decisions (some taxa get a Linnéan category, others not) and gives recent forms more importance than fossil ones, a decision without rational justification. Its use is restricted to side branches of lineages that lead to recent forms, which get a name, while the sister group usually remains unnamed (Willmann 1987).

When the possibility to abandon categories exists, the hierarchy of taxa should be shown with dendrograms or in a subordination (sequential) list. The phylogenetic system of Plathelminthes (flatworms) is described with the following subordination list (simplified; after Ehlers 1985):

Plathelminthes
  Catenulida
  Euplathelminthes
    Acoelomorpha
    Rhabditophora
      Polycladida
      Neoophora

Taxa in the same column belong to the less indented taxon written above. When adelphotaxa (sister groups) are recognized, only two names appear in one column. An unresolved polytomy is indicated when more than two names are listed per column.

12.3.2 Phylogenetic nomenclature

A reform of the traditional nomenclature rules is currently under debate. A set of rules published as "PhyloCode" considers several of the points discussed in the previous chapter (among others: Cantino et al. 1999, see http://www.ohiou.edu/phylocode/). This initiative has also been heavily criticized (e.g., Benton 2000). Major rules of the proposed phylogenetic nomenclature are:

- It is not necessary that all clades be named.
- The system of nomenclature described in this code **is rankless**. (This is the wording in PhyloCode, but what is meant is that there are no mandatory categories to characterize taxa of different inclusiveness.)
- In order for a name to be established under the PhyloCode, the name and other required information must be submitted to the PhyloCode **registration database**. The requirements are: correct publication, correct phylogenetic definition of the clade with the help of specifiers, correct naming (new name or conversion of pre-existing name), registration under the auspices of the Society for Phylogenetic Nomenclature in the PhyloCode registration database.
- In order to be established, the name of a clade must consist of a **single word** and begin with a capital letter. This is also demanded for species names.
- In order to be established, a clade name must be provided with a **phylogenetic definition,** written in English or Latin, linking it explicitly with a particular clade.
- Examples of phylogenetic definitions are **node-based (**naming two members of the clade to define the common ancestor), **stem-based** (naming the sister group to define a pan-monophylum), and **apomorphy-based** definitions.
- **Specifiers** are species, specimens, or synapomorphies cited in a phylogenetic definition of a name as reference points that serve to specify the clade to which the name applies. More than two specifiers may be used.
- Nomenclatural uniqueness is achieved through **precedence,** the order of preference among established names.
- Homonyms may refer to the same taxon under some phylogenetic hypotheses but to different taxa under other hypotheses.

The stability obtained with these rules concerns the spelling of names and the phylogenetic definition of taxa, however, the contents of a named taxon are not stable. The proposed nomenclature has some weak points: (1) It is highly improbable that the scientific community will give up binominal species names. (2) The number of taxon names would increase dramatically because names are linked with specific definitions and cladograms (e.g., if a clade is defined by the position of its sister group, a removal of the sister group or of some other specifier due to a change in phylogenetic hypotheses will require renaming of the same taxon). (3) All valid names that are currently being used (about 3 million) have to be defined and registered, and for decades taxonomists will fight about the correct definitions for established names. (4) Many will not accept that a small committee will play the role of a nomenclature police and take decisions for each proposed taxon name.

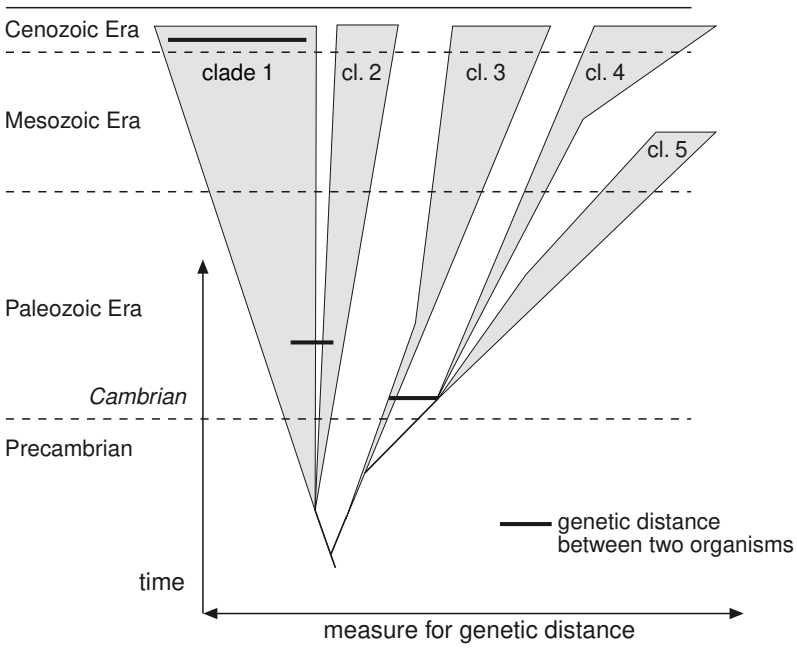


Fig. 183. Classification is independent from true genetic divergence (model, see text).

12.4 Artifacts of formal classification

The formal classification, especially the assignment of categories cannot be used as the basis for statements about genetic distances or divergence times because it is not deduced from evolutionary branch lengths (ch. 3.7). Absurd statements may be the consequence if this is ignored.

Example: the discovery of numerous Cambrian fossils which are classified as stem lineage representatives of recent taxa has often be taken as evidence for a rapid radiation in a comparatively short period of time. For this radiation the term "**Cambrian explosion**" became popular. One of the arguments used by some paleontologists in favour of the occurrence of an extremely fast evolutionary rate and diversification is that most of the animal phyla originated in the Cambrian whereas later no further phyla evolved (e.g., Ohno 1997). This statement is futile and an artifact of classification (Fig. 183): the genetic divergence between a Cambrian stem lineage representative of Echinodermata and one of Chordata (both taxa with the category "phylum") was probably smaller than the difference between two mammals living today (e.g., representing two taxa of the category "family"). Only evidence for an unusually fast genetic divergence between Cambrian fossil species and proving that these taxa really evolved in the Cambrian and had no older last common ancestor can prove the existence of a "Cambrian explosion". Using genetic distances as criterion for the distinction of "phyla", one would distinguish in the present fauna and flora substantially more "phyla" than in the Cambrian.

12.5 Taxonomy

Taxonomy is the science of the description and correct classification of organisms, essential to inventory life forms. Theoretically, the terms "taxonomy" and "systematics" could be synonyms. In practice, however, differences in usage are obvious. A systematist and a taxonomist can conduct different analyses. Systematists search for the phylogenetic system, but they do not necessarily have to acquire special knowledge on the distinction, validity of proper names, and the numbers of known species. Many systematists study the phylogeny of supraspecific taxa but are not able to identify a new species. This, however, can be done by the specialized taxonomist, who knows the rules of nomenclature and how to describe species. The systematist can, but must not necessarily know the rules of taxonomy. Contrary, the taxonomist should know the logics of phylogenetic systematics in order to be able to systematize new species correctly. In practice, however, it is also possible to describe species without knowledge of the theory of phylogenetics. Scientists proceeding this way are taxonomists, but not systematists.

12.6 Evolutionary taxonomy

The attempt to classify organisms according to their relationships *and* organismic complexity or similarity is called **evolutionary taxonomy** (see Mayr 1981). This has the consequence that many taxa are paraphyletic: when separating from the Amniota (Tetrapoda excluding amphibians) those groups that have a higher physiologic performance, namely the Mammalia and Aves, the rest remains as paraphyletic "Reptilia". Paraphyletic groups can be defined *ad hoc* and are not acceptable in phylogenetic systematics. The species *Homo sapiens* could be delimited as a single taxon from other hominids because humans view their capacities as something special (Fig. 184). One would have to distinguish a taxon "apes" from the taxon "humans", a view popular in past centuries. However, this classification does not mirror phylogeny and the historical sequence of speciation events, and therefore evolutionary taxonomy did not gain acceptance.

The procedure necessary for a classification according to principles of evolutionary taxonomy is briefly explained although it does not serve the aims of phylogenetics:

- A character analysis is performed to identify apomorphies, plesiomorphies, and convergences.
- Hypotheses on sistergroup relationships are justified with synapomorphies.
- The number of autapomorphies found in sistergroups is used as a measure for the evolutionary distance to the last common ancestor. The number of new characters is visualized with the length of stem lineages.
- Groups with longer stem lineages get the same rank as the more comprehensive group from which they originate (Fig. 184: separation of *Homo* from extant apes). The "stem group" is therefore paraphyletic.

Birds for example have synapomorphies in common with crocodiles, but they are more different from the architecture (bauplan) of other reptiles than are crocodiles. In the perception of zoologists they have many more autapomorphies. Therefore, in an "evolutionary classification" crocodiles are assigned as reptiles, but birds are not. As a consequence the taxon "Reptilia" becomes paraphyletic.

The disadvantages of this mode of classification are:

- Many taxa are not monophyletic.
- The decision which paraphyletic taxa should be recognized as valid is very subjective and often anthropocentric.

Evolutionary classification as explained here was designed to also consider the degree of genetic divergence between taxa. In doing so the subjective evaluation of visible differences in morphology and physiological performance served as a yardstick. Today, molecular data are available

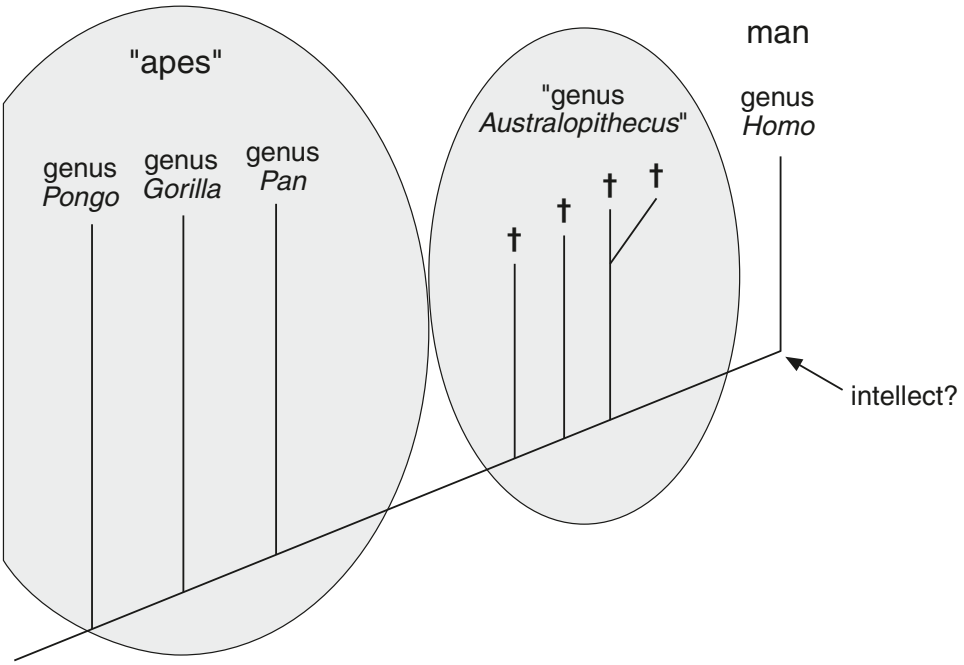


Fig. 184. Separation of humans with the character "high intellectual performance" from other apes: the remaining taxa (names in quotation marks) are paraphyletic.

which allow an objective quantification of genetic divergence, and this information can be used without having to circumvent the rules of phylogenetics.

13. General laws of phylogenetic systematics

There are some generally valid laws of phylogenetic systematics which are independent of the class of characters or the group of organisms or the methods used for reconstruction. They are valid for comparative morphology as well as for molecular systematics.

- To be based on intersubjectively comparable events of nature, a classification of organisms must in the end refer to divergence processes observed between populations (ch. 2.3). (Remember: divergence processes increase the average genetic distance between populations of organisms and can lead to an irreversible separation (= speciation).)
- To build a classification of organisms that mirrors aspects of phylogeny, all distinguished classes of organisms must be monophyla.
- The identification of homologies (ancestral characters or stemline processes) is necessary, but not sufficient to justify a hypothesis of monophyly.
- A hypothesis of monophyly can only be established with apomorphies (ch. 1.3.7, ch. 4). A sistergroup relationship can only be substantiated with synapomorphies. Independent of whether discrete or quantitative characters (genetic distances) are used, evidence has to be presented that identities shared by species are neither plesiomorphies nor chance similarities or convergences.
- The identification of an apomorphy of high probability of homology always implies a hypothesis of monophyly.
- Apomorphies which are used to substantiate a hypothesis of monophyly have to be characters of high probability of homology (ch. 5.1).
- When characters of low probability of homology are used, more complex character patterns can be found by combining simpler characters (ch. 5.1.1, Fig. 88, ch. 6.5).
- Probability of homology depends in a phenomenological approach on the complexity of a character (ch. 5.1.1).
- Information content and probability of homology of a character are the same.
- Statements on homology are always hypotheses (ch. 1.3.7).
- Statements on monophyly are always hypotheses. Monophyla are constructs (ch. 2.6).
- Statements on phylogeny are always hypotheses. Phylogenetic trees are constructs.
- When a dataset contains an insufficient sample of species, hypotheses of monophyly can be mistakenly supported by symplesiomorphies (ch. 6.3.3).
- Methods of tree reconstruction show false monophyla which in reality are para- and polyphyletic groups when (a) unrecognised symplesiomorphies dominate numerically over competing synapomorphies and/or when (b) analogies dominate numerically over competing synapomorphies (errors caused by "long branches", ch. 6.3.2, 6.3.3).
- Fast evolving characters are suited for phylogenetic analyses considering short periods of time, for example, for the comparison of populations or sister species. However, such characters also get noisy rapidly and do not carry any phylogenetic information after longer periods of time. When characters evolve slowly they are suited for the analysis of geologically older speciation events, but they show fewer mutations and a larger number of such characters might be required to get enough information.
- Hypotheses of homology which are required to support hypotheses of monophyly have to be worked out *prior* to the reconstruction of phylogenetic trees. The identification of homologies *after* tree construction can lead to circular reasoning when the same character is used as an argument to support a clade.

14. Appendix: methods and terms

This chapter contains detailed descriptions of some methods, so that if required the reader can obtain a deeper understanding of the selected methods. In view of the wealth of procedures suggested for data analysis in the literature, of which many have not yet proven their worth, and due to the complexity of some mathematical deductions, only a selection is presented here.

14.1 Models of sequence evolution

Within the scope of this text, it can be explained with some clear and simple examples what practical use models have in phylogenetics and how model parameters are estimated (see also general comments in ch. 8.1). It is especially important that the systematist familiarizes himself/herself with the basic principles. The formal descriptions of algorithms, which can be implemented in computer programs, can be left to mathematicians or bioinformaticists. The more complex the models, the more complicated are the corresponding formulae. More important for the user is the identification of implicit assumptions required by specific models. An overview of such assumptions has already been presented in ch. 8.1.

14.1.1 Jukes-Cantor (JC) model

This simple model (Jukes & Cantor 1969) serves the correction of visible distances. This correction is necessary because it must to be assumed that in many cases the number of different character states visible in pairwise sequence comparisons is lower than the real number of substitutions separating the sequences. The number of invisible multiple substitutions increases with the evolutionary distance (remember the "saturation" phenomenon, Fig. 43).

The visible nucleotide difference in a position of two homologous sequences may be caused by a single substitution (A→C) or by several substitutions (A→T→G→C). In the first case, the visible distance (p) would be the same as the evolutionary distance (d) for this position (p = d = 1), in the second case, however, the real distance is underestimated (p = 1, d = 3). This difference should be corrected, for example with the simple the JC-model. To use the JC-model the following assumptions must be valid:

1) The substitution probability does not change in the course of time.
2) The substitution probability is the same for all sequence positions.
3) Sequence evolution is a stochastic (non-chaotic) process.
4) The substitution process is the same in each direction along the time axis (the model is reversible).
5) In the studied sequences the ratio of the bases A:G:C:T is always 1:1:1:1.
6) The base frequency of a gene remains constant in the course of time.
7) The substitution rate is independent of the bases involved. This means there is only one uniform substitution rate.
8) The variable positions of the alignment are not or only partly saturated.

To assume that there exists only one universal substitution rate means that either all substitutions are completely neutral for selection parameters or that selection pressure is identical for all nucleotides and sequence positions. Under such circumstances the probability that a C, T or G evolved from an A is the same in each case.

Defining p as the visible distance between two sequences (proportion of positions with different character states, ch. 8.2.2) and assuming that the assumptions listed above are valid, the evolutionary distance is obtained with

$$d_{JC} = -\tfrac{3}{4} \ln (1 - \tfrac{4}{3} p)$$

The difference between the recalculated distance d_{JC} and the p-distance increases with increasing values for p (Fig. 168). However, computation is only possible up to values of $p<0.75$ because for negative values the natural logarithm is not defined. This limitation is not a problem in practice, because approaching $p=0.75$ saturation is reached and an analysis is pointless. Alignments with high p-values would be totally "noisy" and in this case it is not possible to estimate the real evolutionary divergence. Even single informative positions within a conserved alignment region can be saturated with substitutions. One would get small p-distances when the number of variable positions is low and the *invariable ones are not excluded* from the distance estimation, which is often forgotten in practice. In this case the assumption of a uniform rate for all sequence positions is not valid and therefore the JC-correction would be misleading.

The significance of this model is understood more easily considering the **derivation of the formula** (e.g., Li & Graur 1991):

If an A is present in a sequence position of a gene, with increasing time of existence the probability decreases that the A remains unchanged. This depends on the probabilities that a C, a G or a T is inserted. In the Jukes-Cantor model this substitution probability α is the same for all nucleotides. Thus α represents the probability that after 1 unit of time (which does not have to be defined) a substitution with a specific nucleotide occurs. The probability W that *no substitution* occurs in a unit of time is:

$$W=1-3\alpha$$

After a second unit of time the probability for the conservation of a nucleotide is even smaller, namely (a) the probability that after the first unit of time the nucleotide is still there ($W=1-3\alpha$) multiplied with the same value, plus (b) the probability that after the first unit of time a substitution occurred (3α or $1-W$), multiplied with the probability (α) that from this state again an A originates. The term then reads:

$$W_{(t=2)}=(1-3\alpha)W+\alpha(1-W)$$

Generally, for the time interval $t+1$ there is:

$$W_{(t+1)}=(1-3\alpha)W_{(t)}+\alpha(1-W_{(t)})=W_{(t)}-4\alpha W_{(t)}+\alpha$$

For the difference $W_{(t+1)}-W_{(t)}$ in a unit of time one thus gets $-4\alpha W_{(t)}+\alpha$, or

$$\frac{dW_{(t)}}{dt}=-4\alpha\cdot W_t+\alpha$$

The solution of this differential equation for the probability that the original nucleotide i is preserved ($i\Rightarrow i$) is:

$$W_{ii(t)}=\frac{1}{4}+\frac{3}{4}e^{-4\alpha t}$$

For the case that the nucleotide originally was not present one gets in a similar way:

$$W_{ji(t)}=\frac{1}{4}-\frac{1}{4}e^{-4\alpha t}$$

Both formulas approach the value ¼ with increasing time, which means that independent of the starting point, the probability for the occurrence of a specific nucleotide at time $t=\infty$ is always ¼. This is the same as a random distribution of nucleotides with no bias in base frequency.

Considering that *two* sequences have the same nucleotide at a specific site the probability is

$$W_{ii(t)}=\frac{1}{4}+\frac{3}{4}e^{-8\alpha t}$$

because the time interval is now $2t$. If we are interested in the *difference* between these sequences, one minus $W_{ii(t)}$ is the probability that a change occurs:

$$W_{ij(t)}=\frac{3}{4}\left(1-e^{-8\alpha t}\right)$$

Since all substitutions are treated equally in the Jukes Cantor model, the formula is valid for *any* substitution. In this formula, $W_{ij(t)}$ is the result of a substitution process, in other words, the visible distance between two sequences (equivalent to p in the Jukes Cantor formula). The exponent $8\alpha t$ describes the number of substitutions. Since the substitution probability in one sequence is 3α (see above), the number of substitutions (equivalent to the evolutionary distance d) occurring in two sequences is $2\cdot 3\alpha\cdot t$. Therefore, we can write $\frac{4}{3}d_{JC}$ for $8\alpha t$. Replacing the symbols accordingly, one gets the Jukes Cantor formula.

The application of this model for the transformation of distances is explained in ch. 14.3.2.

The Jukes-Cantor correction corresponds to a specific Poisson correction. The latter has the general formula $d = \ln q$, d being the evolutionary distance and q the visible portion of unchanged positions.

The assumptions of the Jukes-Cantor model can be tested. For example, the base frequencies are easily computed. If among 10,000 nucleotides of an alignment the "A" occurs 3000 times, the observed frequency is 0.3 instead of the expected 0.25.

14.1.2 Tajima-Nei-(TjN-)model

In addition to the assumptions of the Jukes-Cantor model it is considered that the base frequency does not have to be 1:1:1:1 (Tajima & Nei 1984). One has to take into account that in the case of a substitution with a specific base which is more frequent in the alignment than other ones, chance similarities shared by two sequences will occur with higher probability than when the new base is less frequent. Therefore the frequency of each base in the alignment has to be ascertained. As a result the model has apparently four different substitution rates, one for each base originating anew (A, G, C or T). So for distance corrections with the Tajima-Nei model there is:

$$d_{TjN} = -b \ln (1 - p/b)$$

In this formula $\boldsymbol{p}$ is the visible (uncorrected) distance and $\boldsymbol{b}$ is a parameter that depends on the base frequencies $\boldsymbol{q_i}$:

$$b = 1 - (q_a^2 + q_g^2 + q_c^2 + q_t^2)$$

Note that with equal frequencies $\boldsymbol{q_i} = 0.25$ for all bases the value b becomes 0.75. In this case d_{TjN} corresponds to the Jukes-Cantor formula (see also ch. 14.3. 2).

The values for $\boldsymbol{q_i}$ are obtained with $q_i = (\Sigma N_i)/2N$ when distance corrections are performed for the comparison of two sequences, whereby $\boldsymbol{N_i}$ is the number of positions with nucleotide i counted in both sequences, and N is the length of the alignment. Thus in this case the average for two sequences is used as base frequency. For other methods the average base frequency is calculated for the complete alignment. It has to be noted that the base composition is by no means calculated from reconstructed ground patterns, instead only terminal sequences are compared and the assumption is implied that the ancestral sequence of the last common ancestor of two species had a base composition corresponding to the average of the terminal species. In many cases this condition will not be realistic.

14.1.3 Kimura's two-parameter-Model (K2P)

Kimura (1980) pointed out that when comparing sequences of closely related species, transitions occur more frequently than transversions, although there exist more base combinations which correspond to transversions (Fig. 42). The reason for this discrepancy is different probabilities for the change of the chemical class of bases due to cell biological processes. Selection pressure is probably higher against transversions, because these change the chemical class of the nucleotides (purine ↔ pyrimidine). Because transversions are less frequent, they can be traced over longer periods of time than transitions. This is so because also multiple substitutions occur less frequently. If $\boldsymbol{U_t}$ is the probability that a nucleotide remains unchanged in a position at the point in time t, and if $\boldsymbol{S_t}$ or $\boldsymbol{V_t}$ are the probabilities that a nucleotide originated from a transition ($\boldsymbol{S_t}$) or a transversion ($\boldsymbol{V_t}$), then due to the ratio of possible substitutions (compare Fig. 42; for transversions there are twice as many possibilities) we get

$$U_t + S_t + 2V_t = 1$$

It can be shown in a similar way as for the Jukes-Cantor model (ch. 14.1.1), that with the assumptions of the K2P model the probability for transitions is

$$S_t = \frac{1}{4} + \frac{1}{4} e^{-4\beta \cdot t} - \frac{1}{2} e^{-2(\alpha+\beta)t}$$

and for a transversion is

$$V_t = \frac{1}{4} - \frac{1}{4} e^{-4\beta \cdot t}$$

Here α represents the substitution rate for transitions, β the rate for transversions.

Differences of substitution probabilities S_t and V_t are based on mutation and selection processes which cannot be described with "dice statistics". These differences must be estimated with empirical observations and can be recognized when the number of visible transitions N_s and transversions N_v are counted. Using the K2P model for distance methods two sequences are compared to each other and the sequence differences which are caused by (at least one) transition or by a transversion are counted. The proportion of transitions ($P = N_s/N$) and of transversions ($Q = N_v/N$) of an alignment with N positions is considered for the distance correction. A substitution rate is not computed in this way, but one gets a distance measure which is (contrary to the visible p-distance) corrected for multiple substitutions and chance similarities if the model describes the historical processes correctly. The formula for distance corrections is (Kimura 1980):

$$d_{K2P} = -\tfrac{1}{2}\ln(1-2P-Q) - \tfrac{1}{4}\ln(1-2Q)$$

For this model the assumptions of the JC-model are required, with the exception that two substitution rates are distinguished. It has to be stressed that variations of selection pressure in time and also rate variations in different regions of a gene are not considered. Tamura (1992) added a variant in which one aspect of unequal base frequencies is considered, namely the GC-content.

When there is reason to assume that transitions are saturated due to multiple substitutions but transitions might be informative, one can use the JC-model instead of the K2P-model and consider only those positions which show transversions.

14.1.4 Tamura-Nei-model (TrN)

Tamura and Nei (1993) suggested a model which allows not only the distinction of rates for transitions and transversions, but it also considers two types of transitions, because substitution rates between purines (A and G) and between pyrimidines (T and C) can be different. Parameters are estimated in a similar way as in the preceding model, but the TrN-model contains 3 parameters (for transversions, transitions of purines, transitions of pyrimidines). The assumptions of the Jukes-Cantor-model (see above) are also valid here, with the exception that there is no universal substitution rate.

14.1.5 Position-dependent variability of substitution rates

Maximum likelihood methods may become inconsistent if the substitution variability is not considered properly in substitution models, potentially giving a strong statistical support for the incorrect tree. To consider among-site rate variation several methods have been proposed. Differences between these methods are more pronounced for long branches, which for example can cause a different dating of divergence events with a molecular clock. Well-known approaches are (Buckley et al. 2001):

(a) determination of the number of invariable sites,
(b) description of rate categories using a gamma distribution model,
(c) a combination of (a) and (b)
(d) determination of the gamma parameter separately for first, second and third codon positions, or for other partitions of presumably different substitution history, assuming constant base frequencies and relative substitution rates per character partition,
(e) determination of single rates for prespecified distinct character classes (different codon positions, different genes; *site-specific rate models*)

These approaches can lead to different estimates of topology, substitution rates, branch lengths and branch support values. Invariable sites and gamma rates models can give more accurate estimates of branch lengths. The gamma models will in general infer longer branches because extreme rates are not suppressed. The site-specific rate models have the disadvantage that they explicitly assume rate homogeneity within each character class.

Gamma distribution

This correction of visible distances considers differences of substitution probabilities in different sequence positions and requires that several discrete classes of substitutions (of the kind transitions/transversions) exist. The simplest idea is that some portion of alignment positions is invariable, the other positions show a universal substitution rate. It can be ascertained empirically that in alignments of gene sequences function-

ally important regions are often invariable. There is, however, no reason for the assumption that the variable positions have a uniform substitution rate. One can expect that there exists a frequency distribution for the variations of rates, and that this distribution differs from gene to gene. The gamma (Γ) -distribution (Fig. 185) has been proposed to describe such frequency distributions of rates (e.g., Uzzell & Corbin 1971, Yang 1994). In principle, this parameter can be added to different models (JC-, K2P-model etc., e.g., Jin & Nei 1990, Tamura & Nei 1993, see also Fig. 160).

This distribution is described with

$$f(r,\alpha,\beta) = \frac{e^{-\beta \cdot r}\, r^{\alpha-1}\, \beta^{\alpha}}{\Gamma(\alpha)}$$

where $\alpha = \bar{r}^2 / V(r)$ and $\beta = \bar{r} / V(r)$, $\bar{r}$ and $V(r)$ being the mean and variance of r respectively. $\Gamma(\alpha)$ is the so-called gamma function. This is defined by

$$\Gamma(\alpha) = \int_0^{\infty} e^{-t}\, t^{\alpha-1}\, dt$$

The parameters α and β determine the shape of the curve. When $\beta = 1/\alpha$ a distribution with the average rate of 1 is obtained. The specific variant of rate distribution is described with the selection of a value for α (see Yang 1996, Swofford et al. 1996, Nei & Kumar 2000). If β is fixed to be equal to α, the mean distribution of r is 1 and the variance $1/\alpha$. The parameter β (scale parameter) stretches the distribution in the r direction and α influences the shape of the curve (Fig. 185). Note that the surface between each curve is always the same. When small values are chosen for α, this implies a high variability of the substitution rate, with many sites evolving slowly and few evolving fast. This approximates a model with a portion of invariant sites. With high values the rates are more uniform and approach models with position-independent rates, the extreme would be a single rate model.

The true value of α (gamma shape parameter) cannot be deduced from observed substitution processes, because the historical evolutionary processes are not accessible to direct observation. As mentioned in ch. 14.6 (maximum likelihood methods), this and other parameters used in models can be optimized in such a way that the dendrograms optimally match the data. Theoretically, each possible topology should be considered in order to vary the model parameters for each topology and to find the optimal match and the shape parameter alpha has to be determined for each topology. This is computationally very expensive, wherefore approximation methods are used to estimate the gamma shape parameter.

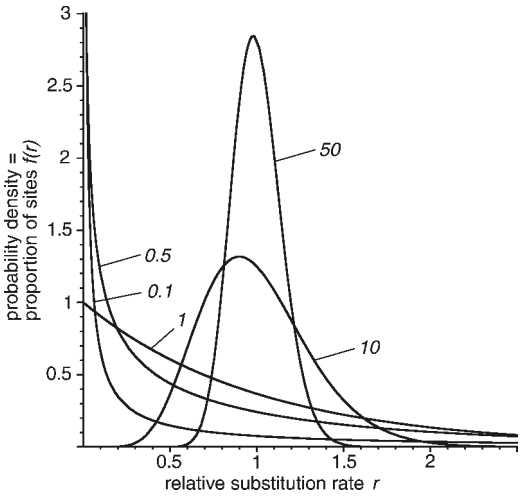


Fig. 185. Examples for different distribution curves of substitution rates r using alpha values between 0.1 and 50. The parameter beta is fixed to be equal to alpha.

Note that using a uniform gamma distribution, the assumption is implied that substitution rates do not change in the course of time along a tree. This assumption may be unrealistic. A related problem is that taxon-specific rate variations are not considered.

Estimation of the position-dependent rate variability

Another method for the consideration of the position-dependent variability in distance methods was proposed by Van de Peer et al. (1993). In principle, the variability of a sequence position is understood to be the frequency of substitutions in a position seen in all stem lineages (branches) of a phylogeny. The variability v_n of a sequence position n is determined as follows:

$$v_n = \frac{s_n \cdot L}{\sum_{i=1}^{L} s_i}$$

Here L represents the length of the alignment, s_n the substitution probability of the position n; s_i is the substitution probability of the position i. For simplicity, the Jukes-Cantor model (ch. 14.1.1) is used for the correction of visible distances. The probability W_n that a substitution is present at a specific sequence position n when comparing two sequences depends on the variability v_n of the position and on the evolutionary distance d:

$$W_n = \frac{3}{4}\left[1 - e^{-\frac{4}{3}v_n d}\right]$$

The variability v_n is determined empirically for a group of positions with similar variability. For this purpose uncorrected pairwise distances of a distance matrix are divided into arbitrarily selected distance intervals (e.g., in steps of 0.005). For all sequence pairs which fall into one distance interval, the percentage of sequence pairs which show a substitution is determined for each position. For example, 135 sequence pairs may show distance values between 0.280 and 0.285, and of these 22 % show a substitution in position 140 (portion of observed substitutions). The portion of observed substitutions can be plotted in a diagram against the distance *for each sequence position*. A point in the diagram for position 140 thus would have the coordinates 0.283 (for the interval 0.280-0.285) and 0.22 (for the portion of observed substitutions). For this group of dots a curve has to be reconstructed with non linear regression which corresponds to the above formula for W_n. The angle of the curve at the origin corresponds to the value v_n of the considered sequence position.

To perform a distance correction considering the specific variability of positions, the values v_n are classified into position groups s of similar variability, each having an average variability v_s. For these positions, f_s is the portion of positions which show a substitution in pairwise sequence comparisons. For the position group s of a sequence pair one then obtains the corrected distance d_s:

$$d_s = -\frac{3}{4}\frac{1}{v_s}\ln\left(1 - \frac{4}{3}f_s\right)$$

The total distance between a sequence pair of the alignment length L is determined by the sum of the distances between the position groups s, each of which is composed of the number of positions L_s.

$$d = \sum\nolimits^{s} \frac{L_s}{L} d_s$$

After each modification of an alignment this calculation has to be performed anew. The method depends on the assumptions of the evolutionary model (here the Jukes-Cantor model).

14.1.6 Log-det distance transformation

The models described above require that the substitution probabilities of specific nucleotides do not change in the course of time, they use *stationary substitution probabilities*. The log-det transformation does not require this unrealistic restriction (Steel 1994, Lockhart et al. 1994, Waddell 1995), and it also allows a shift of base frequencies (G:A:T:C). The log-det transformation is used for distance methods. It relies on the following axiomatic assumptions:

- sequence positions evolve independently from each other and
- substitution rates for a specific type of substitution (e.g., C-A) are constant for all positions of a sequence. Corrections for rate variations at different positions (see gamma distribution, ch. 14.1.5) are not possible.

However, the following parameters are free to vary

- the base frequency,
- the substitution rate in different species and
- the substitution rate at different times.

A basis for distance corrections in pairwise sequence comparisons is a frequency matrix for positions with specific base pairs, whereby all theoretically possible base pairings are considered. The values are obtained by comparing two aligned sequences X and Y.

$$F_{XY} = \begin{pmatrix} f_{AA} & f_{AC} & f_{AG} & f_{AT} \\ f_{CA} & f_{CC} & f_{CG} & f_{CT} \\ f_{GA} & f_{GC} & f_{GG} & f_{GT} \\ f_{TA} & f_{TC} & f_{TG} & f_{TT} \end{pmatrix}$$

In this matrix f_{ij} is the portion n_{ij}/N of base pairs ij in the alignment of the length N. The basic form

of a log-det distance between the sequences X and Y is

$$d_{xy} = -\ln(\det F_{xy}),$$

whereby "det" is the determinant of the matrix. Distances calculated this way are additive (see ch. 14.3.3), but they do not allow an estimation of the number of substitutions that occurred. Assuming that rates do not change with time, d_{xy} can be transformed in such a way that distances are proportional to an "evolutionary distance" (Lockhart et al. 1994). Due to the relative small number of implied axiomatic assumptions this distance transformation is superior to many other models.

A determinant of n^{th} order is the number $\boldsymbol{D}$ which results from the n·n (4×4 in the example above) elements f_{ij} belonging to a matrix as follows:

$$D = \sum (-1)^k f_{1\alpha} f_{2\beta} f_{3\chi} \ \dots \ f_{n\omega}$$

The indices α, β, … , ω pass through all *n!* possible permutations of the numbers 1, 2, … n. Prior to each element of the determinant the sign (+ or –) is determined by the number *k* of inversions in each permutation. For example, the element $f_{13} f_{21} f_{34} f_{42}$ has a negative sign because the arrangement of the second indices shows three inversions ($k=3$) ($3\rightarrow1$, $4\rightarrow2$, and from the first to the last element $3\rightarrow2$) (compare textbooks on mathematics).

In practice it often proves that of all available distance corrections the log-det transformation causes the largest changes of a topology.

14.1.7 Protein coding sequences

Especial models have been developed to consider the different substitution probabilities of amino acids at codon level. It is possible to include in a model nucleotide substitution probabilities that depend on codon changes. Obviously nonsynonymous substitutions should be less frequent than synonymous ones. Since the acceptance probability of a mutated codon is an effect of selection, empirical observations are useful to define codon classes to reduce the number of model parameters, for example, by grouping codons according the polarity and other chemical properties (cysteine: forms disulfide bridges). Stop codons are a separate group. For more details and software recommendations see Schadt et al. (2002).

14.2 Maximum parsimony: the search for the shortest topology

As explained in ch. 6.1.2, it is the aim of MP-methods (maximum parsimony techniques) to find the shortest topology fitting to a given species/character matrix when the optimality criterion is parsimony. Tree "length" is defined as the sum of all character changes found on a topology (compare Fig. 123). A character change is either the occurrence of a new character which did not exist before; or the change of a state in a frame homology (see also the terms "detail homology" and "frame homology": ch. 4.2.2). It can be shown that the search for the "shortest" topology cannot be solved analytically by directed computation with "effective" algorithms (Graham & Foulds 1982). Rather the following steps have to be performed:

– Construction of all topologies which can be produced by combinations of all terminal taxa (ch. 14.2.1), or, if datasets and the number of alternative topologies are too large, heuristic search for topologies (explanation follows below).
– Calculation of the length of each topology considering the selected parsimony criterion (ch. 6.1.2).
– Selection of the shortest topology, or construction of a consensus topology (ch. 3.3) when different topologies show the same shortest length.
– Performance of tests to estimate the probability of recuperating the shortest topology (ch. 6.1.9.2).

Regarding the weights of potential apomorphies or of character transformations formally as sum of single steps (e.g., weight "5" = 5 individual steps), then the **length of the topology** is always the sum of the individual steps on each branch of a topology. This value is independent of the chosen parsimony criterion (see Fig. 127).

Formally, the cladistic parsimony criterion can be expressed as follows (Swofford et al. 1996):

$$L_{(T)} = \sum_{k=1}^{B} \sum_{j=1}^{N} \cdot w_j \cdot diff\,(x_{k'j}, x_{k''j})$$

For a given topology *T,* the length $L_{(T)}$ is the sum of weighted character changes of all N characters (all frame homologies or all columns in a matrix) on all B branches of the topology, whereby for each branch the character states of character j are compared at both nodes or end points (k', k'') limiting the branch. When the character state of these nodes is different, one step multiplied with its weight is counted. The factor w_j is the weight for a change of character j, which has been determined previously (see character weighting, ch. 5.1.2). The weight has to be multiplied with the unweighted number of steps for each character and yields the final number of steps for the character on the branch between k' and k''. When no weighting is intended w_j is 1 for all characters. The total branch length results from the sum of the final number of steps for all character changes on this edge. The total number of steps of a single character is the sum of the weighted character changes on all edges of a given topology. The length of a tree can be calculated from the length of all edges or from the sum of the weighted, topology-specific number of steps of all characters.

In general, the MP method is often recommended because it finds the tree that requires the fewest evolutionary events (in case comparable single events were coded with the same weight) and it allows to assume as little as possible about processes (and corresponding models) of character transformations.

14.2.1 Construction of topologies

Maximum likelihood and parsimony methods require the compilation or at least the consideration of the space of all possible topologies obtained from combinations of terminal taxa. To accomplish this the following methods are used:

1) **Exact search:** an exact search can be performed when the number of taxa and therefore the number of alternative topologies (see ch. 3.4) is small and computation time is short.

   Exhaustive search: this consists of the selection of 3 arbitrarily chosen terminal taxa to construct a first tree, and the successive addition of a further taxon to the previous tree, producing one topology for each connection with one branch of the previous tree. With 4 taxa exactly 3 topologies can be constructed, with 5 taxa already 15 (Fig. 60). As the number of topologies with more than 15-20 taxa exceeds the capacity of many computers, other methods have to be used for large datasets. The complete search is useful to analyse the total tree length distribution (compare ch. 14.9).

   Branch-and-bound search: an exact search is also possible without considering all topologies. After the random selection of the first terminal taxa and addition of further taxa, the tree length (MP-method) or the selected optimality criterion is calculated for each possible topology of the selection. In order to exclude as many topologies as possible and to get to the final result faster, an upper bound for the tree length of a randomly selected topology is calculated. All topologies beyond this length are not considered. This guarantees that more parsimonious topologies will be found, but not longer ones, and the search space is reduced. An acceleration of the calculation can be achieved by determination of a maximal tree length with a preceding heuristic search (see below) (Swofford et al. 1996).

2) **Heuristic search:** Approximation methods accelerate the search drastically, but there is no guarantee that the optimal solution is found (Fig. 186). To reduce the tree space that has to be explored, after each addition of a taxon those topologies which are longer than others with the same selection of taxa are not considered further. In the most simple case one continues to work with the shortest topology and adds the next randomly selected taxon. Thus only some of the possible paths to the complete topologies are followed, which saves a lot of time. A metaphoric description of this method is the climbing of hills during a foggy day: the optimal step during the uphill walk is the one which brings us closer to the top (whereby the "height" is equivalent to the "length" of a topology after addition of a taxon). When the top is reached, the last taxon

of the data matrix has been added to the topology which is most parsimonious on the path taken. However it might be that we reached a side peak which is lower than the main peak, because the path to the neighbouring peaks was not the steepest one at a certain position of the path and therefore has not been chosen. Therefore **local optima** (optima of different paths) and the **global optimum** (optimum for the total dataset) must be distinguished. A heuristic search can end with a local optimum which is not the best solution.

Instead of adding successively one taxon after the other, one can also start with a "star diagram" for all taxa (all terminal taxa linked in one node) and then arbitrarily select one taxon which is placed successively as sister taxon to one of the other taxa (star decomposition method). The topology with the new group that produces the best values with the selected optimality criterion (e.g., the most parsimonious topology in MP methods or shortest distance in distance methods) is retained. An example are clustering methods (ch. 14.3.7) which serve to decompose star diagrams.

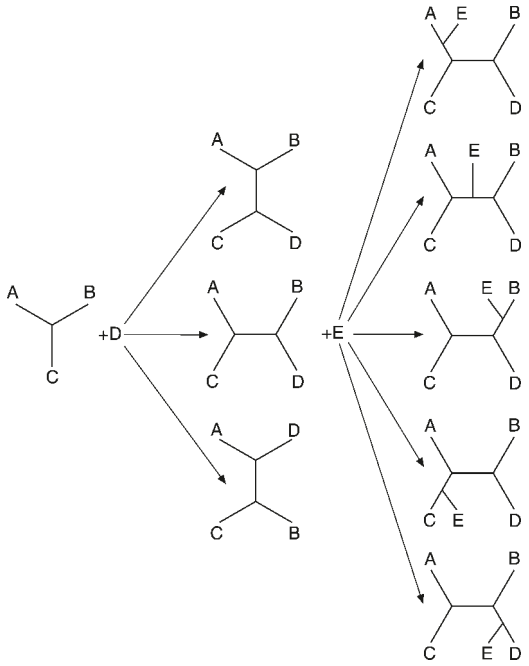


Fig. 186. Example for a heuristic search: stepwise addition of taxa. The taxon chosen next is added to branches of the previously selected topology and of the alternatives only the most parsimonious topology is retained.

Different algorithms for heuristic searches (s. Farris 1970, Saitou & Nei 1987, Swofford 1990) have been implemented in computer programs. Details can be found in the respective handbooks.

3) **Branch swapping.** To improve a heuristic search and to escape local optima it can be attempted to find a modified and more parsimonious topology by shifting of branches. If after the first search a shortest tree is obtained that represents only a local optimum, the chance to find the global optimum is markedly increased by branch swapping. This strategy proves successful in practice. Either two of the four branches joined to an inner branch are interchanged (**nearest neighbour interchange**), or a branch is cut off and added with the cut end to any inner edge of the remaining dendrogram (**subtree pruning**), or the topology is subdivided into two dendrograms at an inner branch (**tree bisection**) and the subtrees are connected again at randomly selected inner branches. For each variant the length is recalculated and compared to the starting topology. Only the topology that is best according to the selected optimality criterion is retained and used as the starting topology for further swapping rounds. If several equally short topologies are found, all these variants have to be considered for further trials. Branch swapping is computationally expensive.

A disadvantage of branch swapping is that in large datasets some sections of a tree might already have an optimal topology, while others need a rearrangement. Random swapping will tear apart optimal sections to improve other parts of the tree, and this increases run times dramatically. To avoid this, new methods that conserve sectors of a tree have been proposed (Goloboff 2001).

When finally several shortest topologies of equal length are obtained, a summary can be illustrated with a consensus tree (ch. 3.3).

4) **Wagner-method.** From a dataset a terminal taxon is selected that shows the greatest similarity with the outgroup or with another randomly selected taxon. Then the next terminal

taxon which produces the smallest number of additional steps (character changes) when added to the topology is searched for in the dataset. For the connection of each additional taxon to the growing topology it is always tested for each branch of the previous tree if the new tree length is shortest. In the end a single topology is obtained. Alternative topologies cannot be detected with the Wagner algorithm and the final tree may be only a local optimum.

14.2.2 Combinatorial weighting

This method is used for weighting of nucleotide substitutions within the framework of cladistic analyses (Wheeler 1990). The basic reasoning is that events occurring more frequently should get a lower weight because they produce more often analogies. To do so, it is tested how often nucleotide substitutions of a sequence position can be seen *in an alignment* while no dendrogram is needed as frame of reference. Therefore this is a form of *a priori* weighting thought to describe the probability of interchange between nucleotides of an alignment position.

For each position it is tested which nucleotides occur simultaneously. The following assumption is required: the more frequently different nucleotides appear associated in columns (positions) of an alignment, the larger is the probability that these nucleotides are correlated by substitutions. The minimal number of substitutions is calculated as (n_k-1), n_k being the number of different nucleotides in a position k. The relative weight for the transformation of one nucleotide into another one is estimated on the basis of the frequency of occurrence of nucleotide pairs in all alignment positions. With four nucleotides there are **six** possible unequal nucleotide pairs ij (Fig. 187), when the polarity of substitutions is neglected (AG, AC, AT, GC, GT, CT).

The association a_{ijk} of the nucleotide pair ij in position k is calculated as

$$a_{ijk} = (n_k - 1) / \binom{n_k}{2}$$

This value will be 1 when only two nucleotides are present, ⅔ with three, ½ with four nucleotides, while $a_{ijk} = a_{jik}$. [Remember: the expression "n_k over 2" means the number of all possible combinations of the n_k elements in groups of two. The solution is obtained with n!/2!(n-2)!]. An "association matrix" with values $\mathbf{A}_{ij}$ for each nucleotide pair is calculated for the complete alignment (Abb. 187):

	A	C	G	T
A	-	2.2	1.5	3.2
C	2.2	-	2.5	1.2
G	1.5	2.5	-	1.5
T	3.2	1.2	1.5	-

Fig. 187. Association matrix (example from Wheeler 1990).

$$A_{ij} = \sum_k a_{ijk}$$

This matrix can be transformed to consider different base frequencies: if the nucleotide C is much more frequent in the alignment than A, one can assume that in the history of the gene the substitution A→C should have been more frequent than C→A. To describe these circumstances the value A_{ij} is divided in the above matrix by the number z_i of positions in which the starting nucleotide i (nucleotide of the left column) occurs. The values are normalized in such a way that the sum of a column is 1. The new matrix with the transformation values T_{ij} becomes asymmetric with this step. The weighting matrix with the weights W_{ij} is obtained with

$$W_{ij} = \left| \ln (T_{ij}) \right|$$

The logarithmic transformation equates the probability of the occurrence of two successive but independent events not with the addition but with the multiplication of the probabilities of individual events. In this way rarer events get a higher weight. The weighting matrix can now be used in MP-methods to weigh in a dendrogram each nucleotide transformation individually.

As with other weighting schemes the length of an edge of a topology is obtained by multiplying the number of character changes $i \rightarrow j$ for each nucleotide pair ij with the weight W_{ij}.

This method requires the following assumptions, which in many cases will be unrealistic:

– The number of observed nucleotide differences of an alignment is a measure of the

historical frequency of substitution events.
- Multiple substitutions do not occur or can be neglected.
- The substitution probability of nucleotide pairs is constant for all sequence regions, for all taxa, and for the whole time.
- Sequence evolution is a stochastic process.

A reasonable suggestion for the consideration of insertions and deletions of individual nucleotides does not exist.

14.2.3 Comparison of MP and ML

The MP method is simple because it does not require detailed models of character evolution that describe the real historical processes. It is, however, possible to include to some degree (without consideration of the effect of different spaces of time) assumptions about probabilities of character transformation (step matrix: Fig. 141). Under these conditions the correct tree is recovered when most of the characters are homologous (low number of convergences and reversals) and when each inner branch is supported by unique character states. Furthermore, when the number of character states is large enough relative to the number of terminal taxa and mutation rates, then MP is statistically consistent for all binary trees (Steel & Penny 2000). (Large numbers of character states are typical for morphological characters, the problem is that the homology of states must be coded correctly.)

Felsenstein (1978) has shown that the MP method becomes **inconsistent** (probability of obtaining the wrong tree increases with sequence length) when too many analogies accumulate in two lineages, which is easily demonstrated for DNA sequences. Dominance of analogies can have several causes: high mutation rates in two independent lineages, but also very low rates in branches connecting faster lineages that cause absence of signal in some inner branches. The latter situation can also produce false groups that share plesiomorphies. (Note that consistency of a method does not help if sequence length can not be increased due to lack of data).

Inconsistency does not occur in the ML method (ch. 8.3, 14.6) when the model used for the reconstruction is a good approximation of the real processes or when the same model was used to generate the data (in simulations). However, ML will not recover the correct tree if the model is not realistic. "Realistic" means that the *effect* of the model is similar to that of the possibly more complicated real processes. For some data (e.g., rare genomic events, evolution of complex morphological characters) models are not available. So, it is not correct to say that ML globally outperforms MP.

Assuming that character evolution can be described with a Poisson model (equivalent to the Jukes-Cantor model in case of DNA data) with the additional condition that rates vary freely from site to site and from branch to branch ("no common mechanism model": Steel & Penny 2000) the maximum likelihood tree(s) (more precisely, the maximum *average* likelihood tree: Steel & Penny 2000) is the same as the maximum parsimony tree(s) (Tuffley & Steel 1997). This is a model that implies that nothing is known about character evolution. However, if some mechanisms are known, probability of finding the correct tree increases when we use this information with the ML method (for further details see Steel & Penny 2000).

14.3 Distance methods

What is a distance? For phylogeny inference different distance concepts are used. For example, a most parsimonious tree is a topology with the shortest sum of paths between terminal taxa, it is based on the ***Manhattan distance***. The *Manhattan distance* is defined as the distance between two points measured along axes at right angles. In a plane with point 1 at coordinates (x_1, y_1) and point 2 at (x_2, y_2), the *Manhattan distance* is

$$|x_1 - x_2| + |y_1 - y_2|.$$

You can move from point 1 to point 2 either taking first the path from (x_1, y_1) to (x_2, y_1) or from (y_1, x_1) to (y_2, x_1), but when finally arriving at point 2 the total distance is the same. It is like searching

the shortest way moving around blocks in Manhattan: there can exist many equally short paths. To estimate the necessary least effort, we are interested in the distance, not in the path. In maximum parsimony analyses we search for the topology with the lowest number of steps (character state changes).

The so-called "distance methods" require a different approach: they are based on pairwise distances.

14.3.1 Definition of the Hamming distance

The Hamming distance is a count of the absolute visible difference between two sequences. For two sequences s and t of the length N and with the respective sequence elements (e.g., nucleotides) $\boldsymbol{s_i}$ and $\boldsymbol{t_i}$ of position $\boldsymbol{i}$ of the alignment $\boldsymbol{S}$, the Hamming distance $\boldsymbol{d_H}$ is defined as

$$d_H(s,t) = \#\{i \mid s_i \neq t_i,\ 1 \leq i \leq N\}$$

If the aligned sequences are of different length, the insertions in sequence s will be opposed by gaps in the other sequences. Deletions are also marked with gaps. It cannot be seen in the alignment whether gaps are caused by insertions or deletions (this is why gaps are also called "indels" (*in*sertions or *del*etions)). Alignment techniques (s. ch. 5.2.2.1) should be used to determine the positional homology. If the latter is correct, then gaps will correspond to real historical events. The Hamming distance for DNA sequences can be interpreted in such a way that a gap is defined as a fifth nucleotide, implying that each inserted or deleted nucleotide corresponds to a substitution event. If this distance is expressed as a proportion of the length of an alignment with the number of positions $\boldsymbol{N}$, one arrives at the $\boldsymbol{p}$ distance: $\boldsymbol{p = d_H/N}$ (see also Dress 1995).

14.3.2 Transformation of distances

The differences counted comparing two sequences do not represent all historical substitutions which occurred since the divergence from the last common ancestor when **multiple substitutions** have modified the same positions. The probability that multiple substitutions occur increases with the divergence time. Therefore the apparent p-distance has to be distinguished from the evolutionary d-distance, as already explained in ch. 8.2. Substitution models are used to correct p-distances. The values for the model parameters (like base frequencies) are usually estimated from pairwise sequence comparisons and not from the whole data matrix, i.e. the parameters can have different values for each sequence pair.

In the Jukes-Cantor model, theoretically the maximal p-distance is $p = \frac{3}{4}$; which corresponds to a random distribution of nucleotides when base frequencies are equal for all nucleotides. The maximal evolutionary distance, however, is theoretically $d = \infty$. High values of p are usually not observed because some positions remain constant due to functional constraints, others show analogous identities which together with multiple substitutions decrease the visible distance in comparison with the true distance.

Therefore two factors have to be considered to estimate the real distance: (a) multiple substitutions and (b) chance similarities which are present even without multiple substitutions and reduce the visible distance.

Distance transformations are required to convert the visible distance into an estimated evolutionary distance using models of sequence evolution. How this is done will be explained in the following using the example of the Jukes-Cantor model which was introduced in ch. 8.1 and 14.1.1 and which is the simplest one.

Jukes-Cantor model

The substitution rate λ tells how many substitutions are to be expected on average for any nucleotide at a sequence position per unit of time (for example per year). Theoretically, λ can have a value between 0 and more than 1, but for empirical data it is usually little more than 0 (e.g., $0.24 \cdot 10^{-9}$ substitutions per year). The rate λ is at the same time the value for the probability that a substitution occurs at one position in the time interval $\boldsymbol{t+1}$ (meaning after one unit of time). Defining $\boldsymbol{w}$ as the probability that no substitution occurs, $w = 1$ means that in this time interval no substitution is to be expected (holds for $\boldsymbol{\lambda = 0}$). The rate is the same for all nucleotides, all substitutions, and all organisms. Considering the two

sequences S_1 and S_2 which diverge since time t, the total number of substitutions is $2\lambda t$ (compare Fig. 163). Assuming that no analogies and no multiple substitutions occur and that the rate is the same in all lineages, then this would also be the number of visible differences (p-distance = d-distance).

After the time interval $t+1$ a nucleotide of a particular position of a sequence will be substituted with a specific probability. Comparing a position of two sequences which show *the same* nucleotide, the probability that a substitution has occurred has to be calculated with 2λ, and the probability that no substitution occurred is correspondingly $w = 1-2\lambda$. The probability that more than one substitution occurred in a position is neglected (λ^2 for two consecutive substitutions) because it is very low for one unit of time. After a substitution one would of course find a visible difference between the two sequences.

If both sequences had two different nucleotides at a specific position, then there exist for each sequence three possible substitutions. One of these three leads to the evolution of an analogy that decreases the visible distance. All other substitutions have no influence on the distance measure. Therefore, for one unit of time the probability that an **analogy** occurs has to be calculated with $2\lambda/3$ ignoring the probability of multiple substitutions (Fig. 156).

A result of these reflections is that the expected portion q of nucleotides of a sequence that are identical to the initial sequence after a unit of time can be calculated as follows, assuming that q_t represents the portion of identities at the beginning (time t) and q_{t+1} at time t+1:

$$q_{t+1} = (1-2\lambda)q_t + \tfrac{2}{3}\lambda(1-q_t) = q_t + \tfrac{2}{3}\lambda - \tfrac{8}{3}\lambda q_t$$

Note: the portion of identities q is a number between 0 and 1 which does not allow statements on sequence length, the number of constant positions, or the number of variable positions! The larger the number of invariable positions, the smaller is the change of the real value of q after a series of substitutions. This consideration is not included for the estimation of distances with the JC-model, because this model assumes a single substitution rates for all nucleotides.

The estimated portion of identities that should be seen comparing two sequences after a unit of time is calculated with the above formula taking the probability that sequence positions remain unchanged $(1-2\lambda)q_t$ plus the probability that analogies occur where previously no identities were present (portion of differences ($1-q_t$) multiplied with $2\lambda/3$). Considering the chance that analogies evolve, it is also taken into account that with **multiple substitutions** an analogy results on average after three substitutions. Thus the above-mentioned distortion of the evolutionary distance which reduces the countable differences is corrected.

Writing dq/dt for an instantaneous rate instead of $q_{t+1}-q_t$ we get:

$$\frac{dq}{dt} = \frac{2\lambda}{3} - \frac{8\lambda}{3}q$$

Under the condition that at t = 0 (start of the divergent evolution) the correspondence between the sequences is q = 1, the solution of this differential equation is:

$$q = 1 - \frac{3}{4}(1 - e^{-8\lambda t/3})$$

For the portion p of visible differences between two sequences there is $p = 1-q$. As the expected number of substitutions d (evolutionary distance) per position is $2\lambda t$ we get:

$$p = 1-(1-\tfrac{3}{4}(1-e^{-4d/3})) \quad \text{or} \quad d_{JC} = -\tfrac{3}{4}\ln(1-\tfrac{4}{3}p)$$

(A slightly different way to explain this formula has been proposed in ch. 14.1.1.)

In these transformations the varying effects of selection on individual gene regions and on different organisms are disregarded. Sequence evolution is treated like a mechanical random process which is constant in time. When in a sequence region multiple substitutions accumulate locally whereas other areas of the same gene are conserved, a correct estimation of the evolutionary distance is not possible with this technique.

Kimura's 2-Parameter Model

As in nature the substitution rate is not the same for all events, the K2P-model (ch. 14.1.3) is more realistic than the JC-model because it considers the conspicuous, selection dependent differences between transition rates α and transversion rates β (ch. 2.7.2.4). Comparing two sequences, altogether 16 different nucleotide pairs can occur. Of these, four are pairs of identical nucleotides (AA, TT, CC, GG), four pairs are transition pairs (AG, GA, TC, CT) and eight are transversion pairs (AT, TA, AC, CA, TG, GT, CG, GC). When the frequency of these three types of pairs is counted, there is $R=1-P-Q$, with R being the portion of homogenous pairs, P the portion of transition pairs and Q the portion of transversion pairs.

In this model the expected average rate of substitutions is composed of the transition rate and the transversion rate. The latter occurs twice as often because, purely statistically, with homogenous base distribution, twice as many transversion pairs than transition pairs are possible (compare Fig. 42). Thus for the total rate there is $\lambda=\alpha+2\beta$ and the evolutionary distance becomes

$$d = 2\lambda t = 2\alpha t + 4\beta t$$

Using this concept, in analogy to the Jukes-Cantor model (see Kimura 1980), a distance transformation correcting for multiple hits and chance similarities can be obtained with

$$d = -\frac{1}{2}\ln\left[\left(1-2P-Q\right)\sqrt{1-2Q}\right]$$

In the Jukes-Cantor model the substitution rate λ is not really calculated, and also with the K2P distance correction no statements on absolute rates are obtained. It is generally assumed that the visible frequency of transition pairs P and transversion pairs Q of an alignment of two sequences are the result of a stochastic process, in which the rates for transitions and transversions are different but constant in time. In the end one applies the Jukes-Cantor model for both of these two types of nucleotide pairs. This implies the assumption that the visible sequence differences are a reliable evidence for the rates of the past. This would be true when these rates were identical for all nucleotides, for all organisms, and at any time. Furthermore it has to be assumed that the sequences are not close to "saturation", because in that case distance estimations are very unreliable.

Distance correction with maximum likelihood parameters

If we assume that pairwise sequence comparisons do not yield good estimates of model parameters, an alternative is to estimate parameters with the ML method and to use these for the fast tree inference with distance methods. Since an ML calculation of a large dataset is very time consuming, one can take a subset of sequences for the ML calculation (Hoyle & Higgs 2003) or use Bayesian methods (ch. 8.4). The fixed rate parameters obtained this way can be used to estimate pairwise distances, but we have to assume in this case that the rate parameters are constant on all branches of the tree. This might be unrealistic.

14.3.3 Additive distances

Ideal distance data for phylogeny inference are additive: they fit exactly on one dendrogram. Distances are additive when they meet the metric "four-point-condition" (Bunemann 1971). When four taxa are chosen arbitrarily from a dendrogram, the tree is additive when for the neighbours A, B and C, D the condition illustrated in Fig. 188 is true.

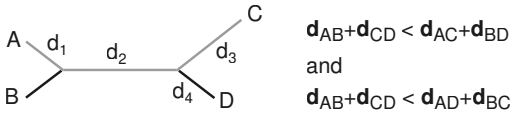


Fig. 188. The four-point-condition holds when distances are additive.

In an additive tree the distance between two taxa is identical to the sum of the length **d** of all edges on the path between the two taxa: $d_{AC}=d_1+d_2+d_3$ and $d_{AD}=d_1+d_2+d_4$.

Genetic distances are additive when substitution rates are identical or different on all edges, but analogies must not be present. Unfortunately,

most real data are not perfectly additive because analogies occur frequently. An analogy can have the effect that the condition $d_{AD} = d_1 + d_2 + d_4$ does not hold any more for the example above (Fig. 189).

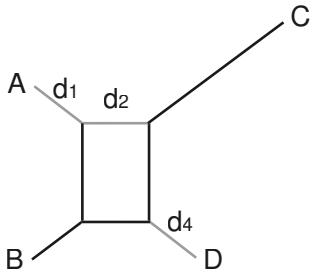


Fig. 189. Distances are not additive when analogies occur. This diagram results when there are some characters shared by A, B compared to C, D as well as characters common to A, C opposed to B, D.

Further rules are:

- Distances between the same terminal taxon are not measurable.
- Distances are independent of the direction in which they are counted ($d_{AB} = d_{BA}$).
- Distances are never negative.
- Distances between neighbours are smaller than the sum of distances between the neighbours and a third taxon (triangle-inequality: $d_{AB} \leq d_{AC} + d_{BC}$).

The triangle-inequality is not met when for example the distances are AB = 2, BC = 1 and AC = 4. In this case for the neighbours A and C (d_{AC}) the path from A via B to C is shorter than the direct path from A to C (in order to understand this you may draw a triangle and enter the branch lengths 2, 1 and 4 on the edges).

14.3.4 Ultrametric distances

Ultrametric distances meet the even stricter "three-point-condition" and fit on a *centrally rooted* dendrogram: the distance between two taxa of an ultrametric tree consists of the sum of the length of the linking edges and in addition the two distances to a third taxon (in the example: A-C and B-C) are equal (Fig. 190):

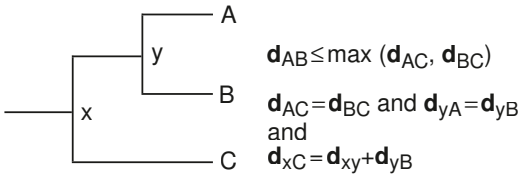


Fig. 190. Ultrametric distances.

Divergence *times* of recent organisms are always ultrametric. When genetic distances are to be ultrametric, substitutions have to follow a perfect and universal molecular clock (Fig. 163). This means that the number of substitutions is proportional to time on all branches of the tree in the same way. In this case it is possible to reconstruct the tree with a simple cluster analysis (UPGMA, ch. 14.3.7) and it is irrelevant which method of distance correction was used to obtain ultrametric data. In reality, however, sequences mostly do not evolve so regularly and usually distance corrections are not so efficient (ch. 2.7.2.4).

14.3.5 Transformation of frequency data to distance data: geometric distances

Genetic distances should be a measure for the time elapsed since two populations started to evolve divergently. The numerically described differences between two populations can be easily summarized as geometric distances. These imply no assumptions about the mode of evolution of the populations considered, which is why the geometric distance possibly has no exact relation to the divergence time, whereas the estimations of genetic distances (see below, ch. 14.3.6) require specific assumptions about processes of character evolution.

Frequency data are data describing the occurrence of allozymes or restriction fragments in individual populations. These data can only be used for phylogenetic analyses when it is guaranteed that the character frequency is characteristic at species or higher taxon level and does not vary much from population to population. Furthermore, a larger number of loci should be considered to obtain a better probability of homology of similarities (due to a higher complexity of character patterns). The transformation of frequency data to distances can be performed with different formulas.

Take a locus ***A*** of the population ***X*** with two alleles present with the frequency x_1 and x_2 ($x_1+x_2=1$). The population ***Y*** has the frequencies y_1 and y_2. Represent each population with a dot in a two dimensional space, the coordinates for point ***X*** being determined on the vertical axis by the value x_1, on the horizontal axis by x_2. Point ***Y*** is defined in the same way. The Euclidean distance between *X* and *Y* then is:

$$d_{XY} = \sqrt{[(x_1-y_1)^2 + (x_2 - y_2)^2]}$$

When ***i*** alleles are present the distance is defined with

$$d_{XY} = \sqrt{[\sum_{j=1}^{i}(x_j - y_j)^2]}$$

Here x_j and y_j are the allele frequencies for allele j in the populations X and Y respectively. This measure does not take into account the probability that with increasing distance multiple substitutions produce analogies. Geometric distances of this kind contain no conceptions about the evolution of gene frequencies, they are only a measure for the similarity of allele frequencies. It remains open which process causes this similarity (for further distance measures see Weir 1996, Swofford et al. 1996).

14.3.6. Nei's genetic distance: allele frequencies, restriction fragments

Take an allele ***i*** of the locus ***A*** present in the populations or species ***X*** and ***Y*** with the frequencies x_i and y_i. If two gametes fuse, x_i^2 is the probability for a chance fusion ***i/i*** (homozygote pairing) in populations of species *X*. When populations ***X*** and ***Y*** mix panmictically (every partner is accepted and individuals are randomly distributed), theoretically gametes with allele ***i*** could meet with the probability $x_i \cdot y_i$. Should the populations have *the same* allele frequency, the probability would be $x_i \cdot y_i = x_i^2$. Nei (1972) defines as a measure for the genetic identity of two populations in the locus A:

$$I = \frac{\sum_i (x_i \cdot y_i)}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}$$

When several loci ***l*** are analysed, allele frequencies of all loci have to be considered, whereby for the populations or species *X* and *Y* there is: $J_x = \Sigma_l \Sigma_i x_i^2$ and $J_{xy} = \Sigma_l \Sigma_i x_i y_i$. The genetic identity of the two populations at the analysed loci is calculated with:

$$I = \frac{J_{xy}}{\sqrt{J_x \cdot J_y}}$$

J_{xy} can be understood as the average probability for a selection of the same allele from two populations by chance alone. J_x and J_y are the probabilities that by random selection within a population (*X* or *Y*) the same allele is found. Nei's genetic distance is defined as:

$$D = -\ln I$$

The value *D*, which theoretically varies between 0 and infinity, is considered to be the estimation of the number of substitutions between two sequences. The concept of the genetic distance according to Nei requires a specific model of evolution of populations: substitutions should always occur randomly and with the same frequency. Therefore, the considered periods of time should be long enough to allow a dominating effect of genetic drift and random mutations on the divergence of the populations. Furthermore, alleles of the common ancestral population should have been in Hardy-Weinberg equilibrium (see Weir 1996). When these conditions are not met because selection is effective, it is a rule for most interspecific comparisons that some loci or some populations evolve faster. For these cases this type of distance estimation should not be applied. Instead, Hillis (1984) recommends the following measure where ***L*** is the total number of the loci:

$$D = -\ln\left[\sum_L \frac{1}{L}\left(\sum x_i y_i \Big/ \sqrt{\sum x_i^2 \sum y_i^2}\right)\right]$$

Working with restriction fragments (**RFLP**-analyses, ch. 5.2.2.4), a simple measure for the similarity of two populations or individuals is the portion of shared fragments

$$F = 2n_{XY} / (n_X + n_Y)$$

where n_x and n_y are the total number of fragments in the populations or individuals *X* or *Y*, while n_{xy}

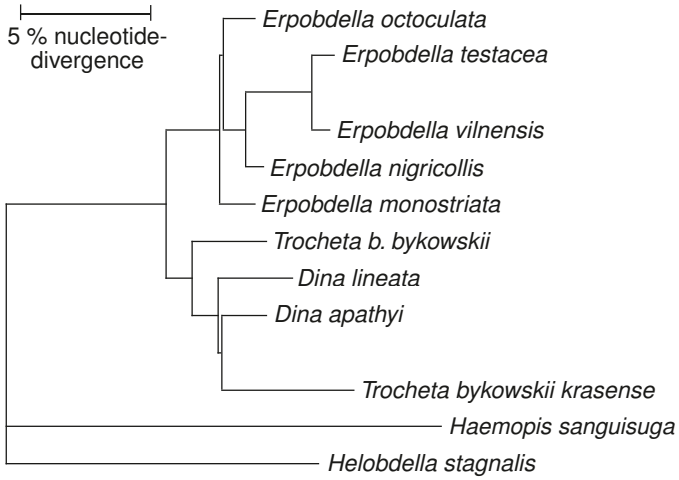


Fig. 191. Relationships between leeches (Hirudinea). Neighbour-joining clustering method, topologies calculated on the basis of RFLP-data from rDNA genes (after Trontelj et al. 1996).

represents the number of shared fragments (Nei & Miller 1990). "Shared fragments" refers to fragments of the same length. The genetic distance between two homologous sequences is obtained with

$$D = -(\ln F)/n,$$

where n is the number of base pairs of the recognition sequence of the restriction enzyme. A dendrogram can be derived from a distance matrix (that results from pairwise sequence comparisons) using the *neighbour-joining* clustering method (Fig. 191 and ch. 14.3.7).

It is recommended to calculate separately the distance for restriction enzymes of different length, because longer enzymes cut with greater probability at homologous loci. Defining k as the number of different enzymes with the length r, and m_k as the average number of cleavage sites of the enzymes of class k (this means $m_k = (n_{Xk} + n_{Yk})/2$), D_k being the genetic distance which has been calculated with the enzymes of class k, then the total distance between the two sequences is calculated with:

$$D = \frac{\sum_k m_k r_k D_k}{\sum_k m_k r_k}$$

However, the individual identification of restriction fragments and character coding for parsimony analyses is to be preferred over the computation of distances, because then probability is higher that homologous substitutions are considered (ch. 5.2.2.4). For distance estimates using RFLP-data, strictly speaking, the following axiomatic assumptions have to be met:

- all nucleotides are equally frequent in the genome or in genes studied.
- The modification of cleavage site composition and number is only caused by base substitutions.
- Substitution rates are the same for all nucleotides and species.
- No restriction fragments are overlooked (which may happen with larger numbers of fragments), and different fragments of the same length must be discerned.

14.3.7 Construction of dendrograms with clustering methods

UPGMA

This abbreviation stands for **u**nweighted **p**air-**g**roup **m**ethod using arithmetic **a**verages. With this method one compares distances between objects (sequences, terminal taxa) pairwise and calculates the distances between most similar pairs to other objects, using arithmetic averages

that represent groups of objects. With this tool one cannot only compare individuals or species, but also (in ecology) samples from different regions, for example, when some similarity index has been calculated for species diversity in samples. The following example explains the procedure when a number ***S*** in the matrix is a measure for similarity between two objects:

species	1	2	3	4	5
1	/				
2	0.05	/			
3	0.04	0.07	/		
4	0.03	0.09	0.2	/	
5	0.14	0.05	0.04	0.04	/

In this matrix species 3 and 4 are the most similar. This pair is the first sistergroup in the growing dendrogram. Now a new matrix is calculated, wherein 3 and 4 are united to a new terminal taxon. The similarity between this pair (3, 4) and species 1 is obtained with

$$S_{(3,4)1} = (S_{1,3} + S_{1,4})/2 = (0.04 + 0.03)/2 = 0.035.$$

This similarity value is entered in a new matrix:

species	1	2	5
1	/		
2	0.05	/	
5	0.14	0.05	/
(3,4)	0.035	0.08	0.04

The next pair is (1, 5). The procedure is continued to calculate all distances in this same way. However, it is already clear that the topology will be ((3,4)(1,5)2).

The UPGMA method requires the assumption that the substitution rate is constant in all lineages up to the terminal taxa, which means that genetic distances are proportional to the divergence time and ultrametric (Fig. 190). When this condition is not met, one often gets wrong results because autapomorphies can increase the distance in one of two sister taxa to such an extent that the sistergroup relationship is not recovered. This clustering method is not recommended because evolutionary processes generally do not run so regularly. UPGMA is nevertheless useful to learn how clustering techniques work.

Neighbour-joining

This method can be used when distances are not ultrametric, a "universal molecular clock" for evolutionary processes is not required (Saitou & Nei 1987, explained in Swofford et al. 1996). At first, pairwise evolutionary distances d_{ij} for terminal pairs of taxa (***i, j***) are estimated (compare ch. 8.2.1) which are subsequently corrected to obtain an average distance between this pair and other taxa. This corrected matrix value M_{ij} is used for the further calculations.

It is possible to define for a species (a terminal taxon) ***i*** a total distance r_i to all other species, ***T*** being the number of terminal species:

$$r_i = \sum_k^T d_{ik}$$

With five species one would get for the first species: $r_1 = d_{12} + d_{13} + d_{14} + d_{15}$ and $d_{11} = 0$. The differences in substitution rates of homologous characters along the stem lineages of two species are compensated to obtain the value M_{ij} which corresponds to an average distance value:

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{T - 2}$$

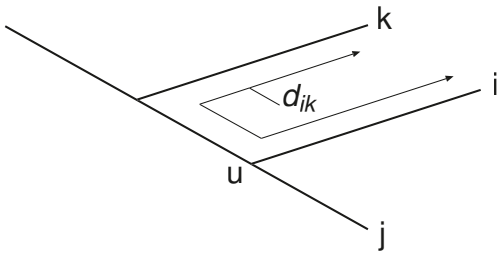


Fig. 192. Diagram for the neighbour-joining method.

In the corrected matrix those species which have the largest negative value of M_{ij} are neighbours. For a neighbouring pair of species identified in this way a ground pattern (basal node) ***u*** is proposed. In a new data matrix the neighbouring species pair (*i, j*) is replaced with the node ***u*** and the distance d_{uk} between this node and each other taxon ***k*** is calculated under consideration of the evolutionary d-distance:

$$d_{uk} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$$

Checking the example in Fig. 192 one finds that the distance d_{uk} is obtained with this formula. This step is repeated until the topology is resolved. When the stem lineage of a terminal taxon is especially long, this has no consequences for the distance estimations, in contrast to the UPGMA technique.

Like other clustering methods, neighbour-joining is sensitive to the order of taxa in the data matrix, which influences the topology of the dendrogram. In order to test the extent of this effect for a given set of data, the computation has to be performed several times with different successions of taxa. The order also influences bootstrap-values which in some cases may be far too high (Farris et al. 1996).

14.3.8 Construction of dendrograms with minimum evolution methods

This method seeks to minimize the sum of branch lengths $\boldsymbol{L}$ in the optimal topology and tree length is computed from pairwise distances. For an unrooted tree with $\boldsymbol{n}$ terminal sequences and $(2n–3)$ branches, the sum of individual branch lengths $\boldsymbol{e_i}$ is

$$L = \sum_{i=1}^{2n-3} e_i$$

To estimate branch lengths $\boldsymbol{e_i}$ we need pairwise evolutionary distances δ_{ij} between terminal taxa $\boldsymbol{i}$ and $\boldsymbol{j}$ and a tree $\boldsymbol{T}$. The distance $\boldsymbol{d_{ij}}$ is the path length between $\boldsymbol{i}$ and $\boldsymbol{j}$ in this tree (tree distance, patristic distance). In the ideal case when the substitution process and the topology are correctly inferred, we should get $\delta_{ij} = \boldsymbol{d_{ij}}$. All *branch lengths* $\boldsymbol{b_k}$ (k being a branch of topology T) are written into a column vector $\boldsymbol{B = (b_k)}$. The topology T is represented with a *topology* matrix $\boldsymbol{A = (a_{(ij)k})}$, in which $(a_{(ij)k}) = 1$ indicates that a branch k of the path between i and j is present, while 0 indicates absence of a branch. For the column vector $\boldsymbol{D = (d_{ij})}$ containing all patristic distances we can also write $\boldsymbol{D = AB}$. This vector is compared with the vector $\Delta = \delta_{ij}$) that contains all pairwise evolutionary distances δ_{ij}:

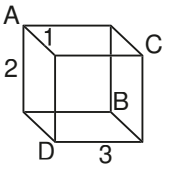
For the ***ordinary least-squares*** approach (Rzhetsky & Nei 1992) the squared Euclidean fit $(D\text{-}\Delta)^T(D\text{-}\Delta)$ between Δ and D is minimized (T stands for matrix transposing (flipping)). To consider differences in variance of δ_{ij}, the least-squares can be weighted with variance estimates for δ_{ij} (weighted least-squares) or with full variance-covariance estimates (generalized least-squares) (for more details see Swofford et al. 1996, Gascuel et al. 2001).

Trees are not constructed directly from a matrix as in clustering methods, and alternative topologies obtained combining terminal taxa have to be tested. Since exhaustive tree searches are time-consuming when the number of sequences is large, heuristic searches, greedy tree constructing methods and branch swapping can be used to find the global optimum. In fact, recently developed algorithms are at least as fast as neighbour joining methods (Desper & Gascuel 2002).

14.4 Construction of networks: split-decomposition

Split-decomposition is a useful tool for the graphical representation of networks (Bandelt & Dress 1992). This approach is based on the following reflection: the topology of a dendrogram is unambiguously determined when the relationship between quartets of taxa of a tree is known (four-point-condition, see ch. 14.3.3). Considering four taxa which differ in three binary informative characters (= characters with only two character states), the largest incompatibility exists when the taxa and their mutual distances can be arranged to form a cube (Fig. 193): in Fig. 193, character 1 establishes the edges parallel to branch 1, character 2 those parallel to branch 2, and character 3 has the equivalent effect. In contrast to a tree, the cube represents exactly all distances between four species. A two-dimensional projection into the plane is obtained by omitting one of the four corners that are not occupied with taxa. This graphic contains also the exact distance ratios. As there are four "empty" corners of the cube, four alternative and equivalent projections can be constructed. The projection is already a small network diagram. Each character produces in the diagram a set of parallel edges, because the three characters of the example are incompatible.

taxa	characters 1	2	3
A	0	0	1
B	0	1	0
C	1	0	0
D	1	1	1



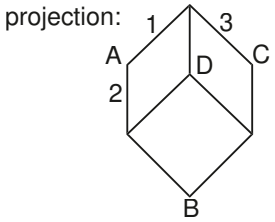


Fig. 193. Matrix and graphs showing the incompatibility of three characters in four taxa (after Bandelt 1994; explained in text).

For four taxa maximally three dichotomous topologies can be constructed (Fig. 194). These topologies are all contained in one cube diagram (Fig. 193): taxon A is a neighbour of taxon B, C, and D, the distance to neighbours is equal in each case. In individual dichotomous diagrams however, distances are unequal (e.g., $d_{AB} < d_{AC}$). In a square with non-empty corners there are two neighbours of equal distance, the distance to the fourth taxon is larger. A cube implies three equidistant neighbours which together are incompatible with a dichotomous dendrogram, a vertex in a square has only two such neighbours. With more than four taxa multi-dimensional cubes (hypercubes) are obtained. Split-decomposition methods permit illustration of sub-graphs (reticular split graphs) derived from hypercubes.

To construct a network, a data matrix is examined character by character. The order in which characters are evaluated is irrelevant. Each binary character produces a split (bipartition) in the set of species: in the graph all species with character state 0 are separated from those with state 1 by a set of parallel edges (Fig. 196). In this way two sub-networks are produced with members showing either the character state 0 or the state 1.

As different characters can fit to the same split, there can be edges with better support than other ones. This fact can be visualized by increasing the length of a branch in relation to the growing number of supporting characters. Symbols for

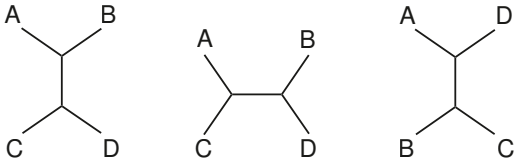


Fig. 194. All possible dichotomous dendrograms for four taxa.

characters which cause the edges can be entered in the diagram and when known, the polarity could be indicated with an arrow. The branch length can also be drawn to scale indicating genetic distances when no discrete characters are used and only a matrix with pairwise distances is available.

For the application of distance measures, "d-splits" for DNA sequences can be described as follows: X is a given set of sequences which can be divided into a family of splits, whereby each split $S=\{A,B\}$ consists of two groups A and B with $A \cup B = X$. Each split is described with the distance between both groups. This distance can be the uncorrected Hamming-distance (ch. 14.3.1) or an evolutionary distance obtained with model assumptions (e.g., Jukes-Cantor distance, compare ch. 14.3.2). For the distances of sequence pairs s, s' of the group A and t, t' of the group B of the split $S=\{A,B\}$ the following condition must be met:

$$d(s,s') + d(t,t') < \max \begin{pmatrix} d(s,t) + d(s',t') \\ d(s,t') + d(s',t) \end{pmatrix}$$

(compare Fig. 195)

For split-decomposition graphs one selects for a quartet of sequences the sums of distances which are not the largest of the three alternatives,

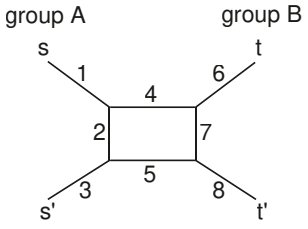


Fig. 195. Split-graph for the sequences s, s', t and t' to explain the condition for distance splits (after Bandelt & Dress 1992). The numbers are indices for the branches (see text).

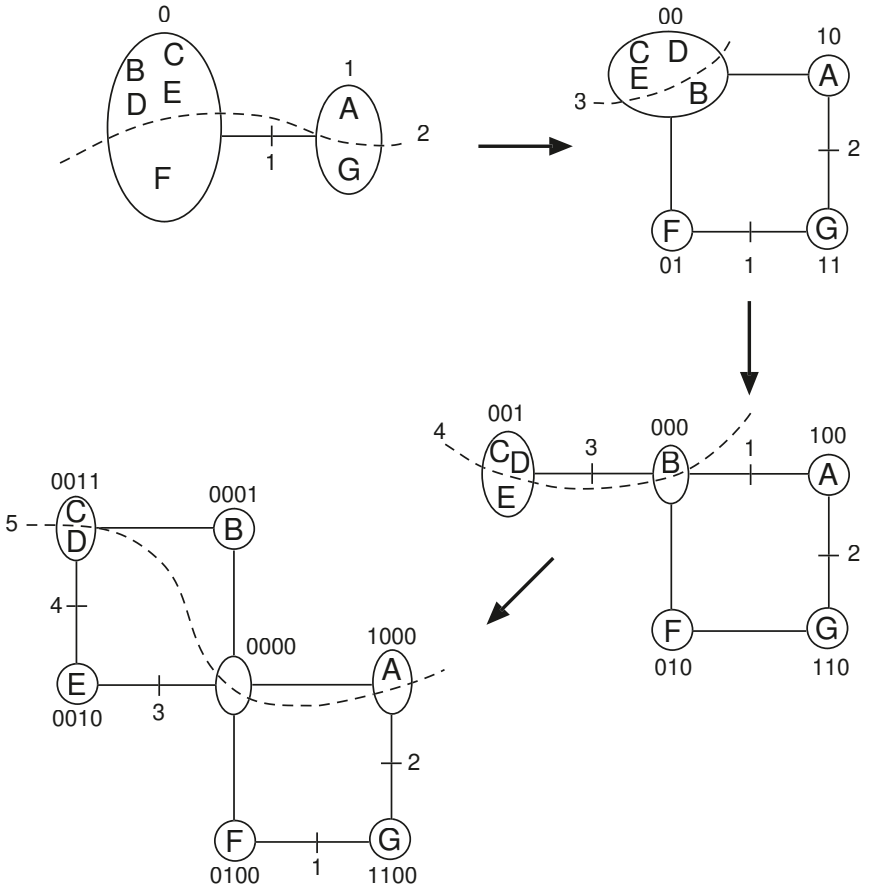


Fig. 196. Examples for the construction of a network (after Bandelt 1994). The example shows the decomposition of a set of species by the first four characters of a fictitious dataset to explain the principle of the split-decomposition method. The letters represent terminal taxa, the numbers above the taxa indicate the states of binary characters, the numbers at the branches are characters which produce a split. Character 3 for example has the state 1 in taxa C, D and E and produces the split {(*C*,*D*,*E*),(*A*,*B*,*G*,*F*)}.

assuming that for the sistergroup relationship which corresponds to the largest sum (e.g., {s,t'} or {s',t} in Fig. 195) there is probably no phylogenetic information present in the data. In Fig. 195 the distance d(s,s') consists of the length of branches 1+2+3. With the sum d(s,s') + d(t,t') the branches 4 and 5 forming the split {*A*,*B*} are excluded from the network. The largest of the three alternative sums excludes the shortest split-supporting edge, which is the one based on the lowest number of character changes. Therefore, with the condition described above the split is excluded which is least supported by characters or distances. Omitting one of the three parallel edges of a cube for each quartet of terminal taxa, the relations between all taxa of a dataset can be illustrated as two-dimensional network (Figs. 150, 196).

For each split an *isolation index* αs which describes the branch length is calculated. For the split {*A*,*B*} this index is

$$\alpha s = \tfrac{1}{2} \min (\max (d(s,t) + d(s',t'), d(s,t') + d(s',t)) - d(s,s') - d(t,t').$$

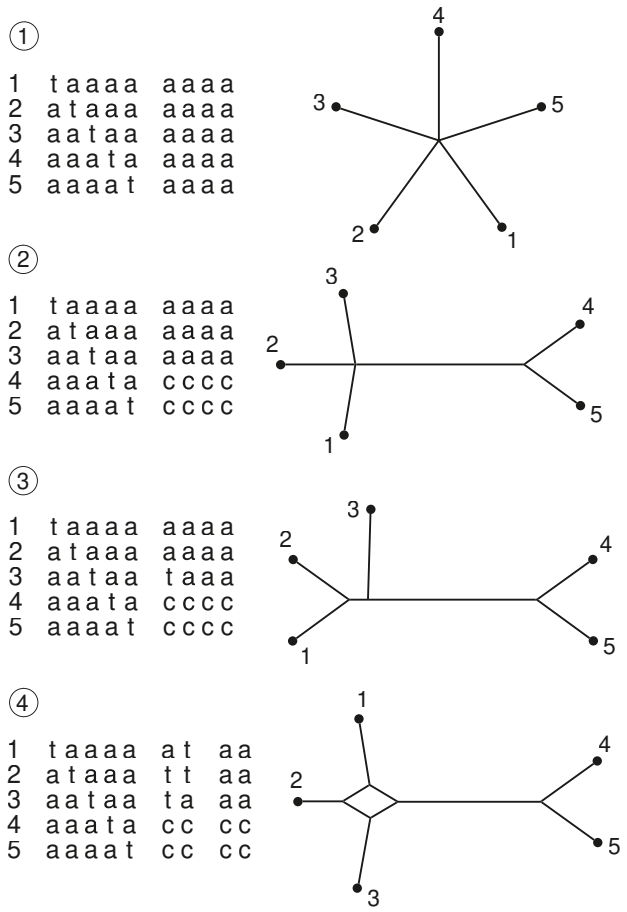


Fig. 197. Examples of split-graphs obtained from artificial, clear-cut datasets. A comparison of the alignment with the respective graph shows that autapomorphies of individual taxa cause the branches to the terminal taxa (case 1), additional binary characters occurring in more than one species form groups of taxa (here only one split, case 2), autapomorphies in split supporting positions decrease distances (case 3), and incompatible split-supporting positions produce networks (case 4).

The isolation index can be considered a measure of support for a split by the given data and is depicted as branch length or indicated as weight. Some branches of the network are not illustrated in two dimensions. Together they add to a rest which can be specified in percent of the total length of all distances of a dataset. The split graph only contains the information that can be depicted two-dimensionally.

In split-decomposition of sequence data, all positions which show for the group *A* the same nucleotide not occurring in group *B* act as **supporting positions** of the split S={*A*,*B*}.

The influence individual characters have in this method can be seen in the example of Fig. 197. Fig. 197 proves that in d-split decomposition only binary characters produce splits, while autapomorphies of terminal taxa increase peripheral branches and noise in the data (autapomorphies, convergences) decreases the resolution of the graph. When discrete characters are used more problems arise: the more taxa the dataset contains, the worse is the resolution of the graph because each autapomorphy in a supporting position decreases the length of inner branches. For large datasets one often gets star-trees which certainly do not represent the available information. Therefore it is better to evaluate the noise *a*

priori (for example with "spectra", ch. 6.5) and to use the number of split supporting positions seen in spectra like a distance measure.

New methods permit a better resolution of split-graphs constructed from a distance matrix. It is possible to use matrices of pairwise evolutionary distances corrected for multiple hits, chance similarities etc. As in cluster analyses (neighbour joining: ch. 14.3.7), the NeighbourNet algorithm (Bryant & Moulton 2002) selects at each iteration a pair of neighbouring nodes and estimates a new composite node, but the original pair is not immediately replaced (a difference to the NJ method). A node must be paired twice, and then the three linked nodes x, y and z are replaced by two linked nodes u and v. Only then the distance matrix is reduced and the next iteration starts. The method generates a "circular split system" (a set of splits that fit to a planar graph in which all terminal nodes can be arranged on a circle) rather than a tree and therefore it is useful to demonstrate contradictions contained in the data. Edge length estimations are performed at the end. The estimation of distances for the distance matrix requires corrections with substitution models if molecular sequence data are used (ch. 14.3), and estimated edge lengths are only correct if the substitution model describes the real substitution process.

There exist alternative definitions for net-like graphs, as for example:

Minimum spanning trees

A **minimum spanning tree** is a tree formed to find the shortest connection between nodes. It has two properties:
- It *spans* the graph – it includes every terminal node (vertex) in the graph and
- it is a *minimum* – the total length (or weight) of all the edges is as low as possible.

The length of the graph is just the sum of all edge lengths or weights. It can look like a network or be more tree-like (Fig. 198). Mathematically, the minimum-spanning tree is not a network because it has no cycles.

Median networks

A median network contains all most parsimonious trees that can be constructed from a dataset. The graphs are constructed from characters that

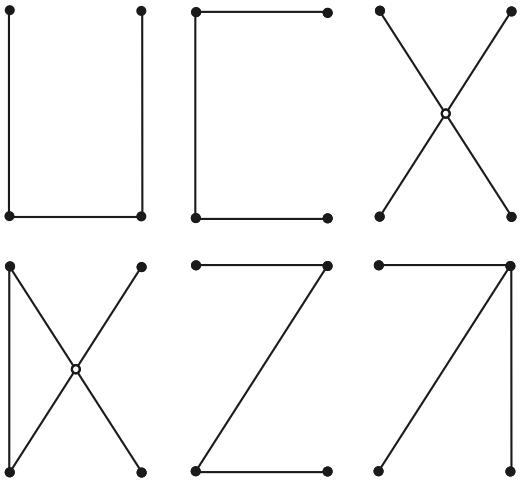


Fig. 198. Examples for alternative spanning trees (there are 16 variations for 4 nodes). The black points represent terminal nodes, the white one is added to search for shorter trees. Mathematically, these graphs are trees as long as they contain no cycles, which means that every pair of nodes is connected by a unique path. The shortest graph is the minimum spanning tree.

are binary. Sequence alignments should be transformed to binary data. Fortunately, working with closely related specimens (e.g., comparing haplotypes) most variable characters will be binary. Constant characters are removed. Ambiguous characters can be omitted while constructing the network and later they are fitted to the graph so that the number of additional links or nodes is minimized. Characters that are compatible are connected by simple branches, whereas incompatible characters added to the tree double parts of the graph and increase the dimensionality of the network (Bandelt et al. 1995, 2000).

All variable sequence positions that support the same split are grouped in one complex "character" which is weighted with the number of supporting positions, the number of "characters" (splits) is $\boldsymbol{k}$. Both sides of a split are coded with 0 or 1 (the numbers have no polarity). Thus each sequence (or haplotype) is represented by a vector of length k, the number of sequences is $\boldsymbol{n}$. Two vectors are linked unambiguously when they differ in a single coordinate.

The three vectors 000, 001, 101 represent for example three sequences, each supporting three different splits (Fig. 199). The vectors can be connected to form a six-cycle or a tree. The tree is the

Splits representing sequences:

sequence	splits u v w
A	0 0 0
B	0 1 1
C	1 0 1

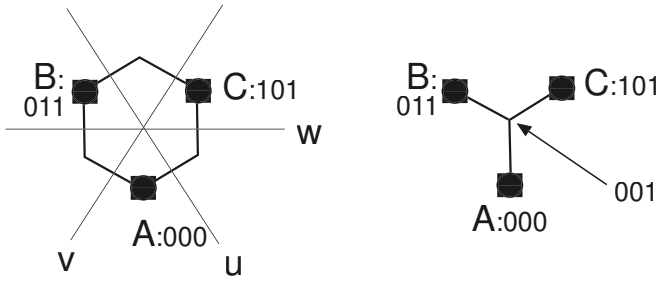


Fig. 199. Two graphs connecting the three vectors 000, 011 and 101.

more parsimonious solution that is obtained constructing the consensus vector (median vector) 001. The distance of this vector from any node is one step. By addition of further triplets obtained by any combination of vectors a median network is constructed. In practice, a *j*th split ("character") is added by splitting the graph that has already been constructed into the groups separated by *j* (Fig. 200). A median network includes the most parsimonious tree for each triplet of terminal nodes.

Median-joining network algorithms can handle larger datasets. To construct a network one starts with a minimum spanning tree and subsequently adds a few median vectors (consensus sequences) of three mutually close sequences at a time to obtain most parsimonious trees for triplets. In contrast to the median network explained above (which can be very complex for large datasets), the median-joining network contains only selected median vectors which have a good chance of representing inner nodes in a most parsimonious tree of the dataset (for more details see Bandelt et al. 1999).

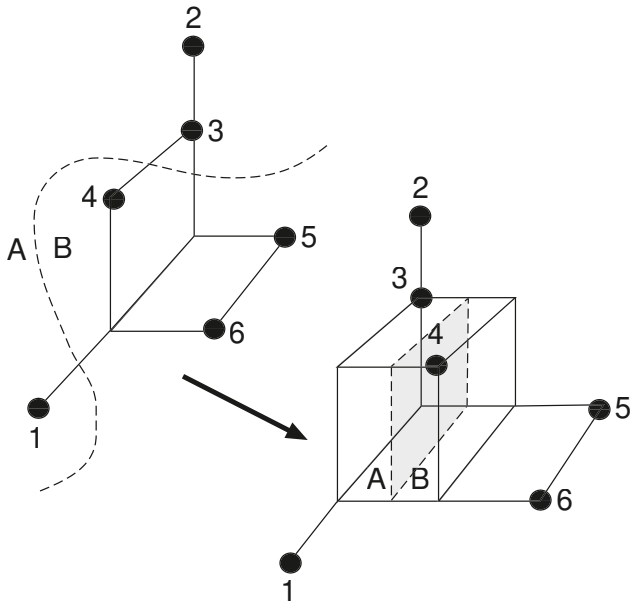


Fig. 200. A step during the construction of a median network: addition of a further character present in the terminal sequences 1-6 that supports the split separating the groups A = {1, 2, 3} and B = {4, 5, 6} (after Bandelt et al. 1995).

14.5 Clique analyses

This method is based on the consideration that binary characters of a set of data can support incompatible splits (compare ch. 6.5.1) and amongst the alternative combinations the dendrogram has to be found that has for its partitions the largest number of supporting characters. Characters which are incompatible with this dendrogram are ignored. Ideas and algorithms were developed by LeQuesne (1969), as well as by Estabrook et al. (1976a, 1976b) and are incorporated in Felsenstein's PHYLIP program package (Felsenstein 1993).

The method requires a data matrix with binary characters. Characters with several states have to be recoded to get binary ones (Figs. 118, 119). A polarity is not required, the dendrograms are unrooted. The following examples elucidate the effects of the method.

The example of Fig. 201 illustrates the principles of the method. A clique is defined as a group of supporting characters for a set of compatible splits. The clique consisting of the characters 1 to 3 plus 8 is slightly smaller than the one for the characters 4 to 7 plus 8. Therefore the characters 1-3 are ignored, and only the split {(B,D), (A,C,E)} is shown in the dendrogram. This is an important difference to split-decomposition which visualizes the extent of conflicting data. Trivial characters (autapomorphies of terminal taxa) have a neutral effect on the analysis.

When there are several cliques of equal size, several equivalent dendrograms are found (Fig. 202).

The method implies the following assumptions:

– The probability of homology of all characters is the same (otherwise the characters would have to be weighted).

data matrix:

	characters							
species	1	2	3	4	5	6	7	8
A	1	1	0	1	1	1	1	0
B	1	1	0	0	0	0	0	0
C	1	0	0	1	1	1	1	0
D	0	0	1	0	0	0	0	1
E	0	0	1	1	1	1	1	0

compatibility matrix:

characters	1	2	3	4	5	6	7	8
1	1	1	1	·	·	·	·	1
2	1	1	1	·	·	·	·	1
3	1	1	1	·	·	·	·	1
4	·	·	·	1	1	1	1	1
5	·	·	·	1	1	1	1	1
6	·	·	·	1	1	1	1	1
7	·	·	·	1	1	1	1	1
8	1	1	1	1	1	1	1	1

characters of the largest clique:
4, 5, 6, 7, 8

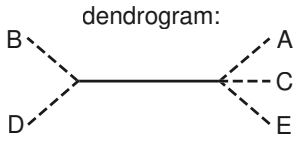


Fig. 201. Example of a clique-analysis. In the data matrix the characters 1-3 support the partitions {(A,B,C),(D,E)} and {(A,B),(C,D,E)}, which are mutually compatible*. In the compatibility matrix therefore a 1 represents the compatibility of these characters. In contrast, the characters 4-7 are not compatible with 1-3, they support the partition {(B,D),(A,C,E)} and are again compatible to each other. This is recorded in the compatibility matrix accordingly. Character 8 is trivial and therefore compatible with all other characters. The group of compatible splits with the largest number of characters is supported by the "clique" of characters 4-8. Only this is used to construct a dendrogram, showing the corresponding partitions (in this example only a non-trivial partition is supported).

* Note that this is not the compatibility of Venn diagrams, but the compatibility of splits that fit to an unrooted tree.

data matrix:

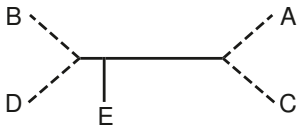
species	characters 1	2	3	4	5	6	7	8	9
A	1	1	0	1	0	1	0	0	0
B	0	0	1	0	0	0	1	1	0
C	1	1	1	0	1	0	0	0	1
D	0	0	0	1	1	0	1	0	0
E	0	0	0	0	1	1	0	1	1

compatibility matrix:

characters	1	2	3	4	5	6	7	8	9
1	1	1	·	·	·	·	1	1	·
2	1	1	·	·	·	·	1	1	·
3	·	·	1	1	·	1	·	·	·
4	·	·	1	1	·	·	·	1	1
5	·	·	·	.	1	·	·	·	1
6	·	·	1	.	·	1	1	·	·
7	1	1	·	.	·	1	1	·	1
8	1	1	·	1	·	·	·	1	·
9	·	·	·	1	1	·	1	·	1

largest cliques:

I) characters 1, 2, 7



II) characters 1, 2, 8

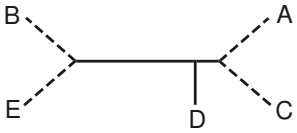


Fig. 202. Example of a clique-analysis for data with two equivalent cliques.

– A marginally higher number of shared states is already interpreted as decisive phylogenetic signal (or to be more precise, a *homology signal*).

When instead of the phenomenological probability of homology the probability of events is evaluated, the following assumptions prevail:

– The probability that convergences or chance similarities evolve is smaller than the probability that apomorphies originate.
– Characters evolve independently of each other.

The method can be used to analyse patterns present in datasets (phenomenological analysis). It does not make explicit assumptions about rates of character evolution. However, for the analysis of morphological characters parsimony methods are preferred, because they allow a better evaluation of the support of putative monophyla considering also the presence of conflicting characters (homoplasies). For the interpretation of molecular data spectral analysis clearly has advantages. In contrast to spectral analyses, the ratio of "signal" to "noise" cannot be described with the clique-method.

14.6 Maximum likelihood methods: analysis of DNA sequences

The goal of these methods is to find amongst all alternative dendrograms which can be constructed from a dataset, the one that explains with highest probability the evolution of the terminal sequences when a defined process of sequence evolution is assumed. This process is represented with a suitable model of sequence evolution that is selected and used for the computation.

The following steps have to be executed for the likelihood estimation:

a) Selection of one of the alternative dendrograms and
b) selection of a model of sequence evolution. Some model parameters can be estimated from the data.
c) Analysis of the probability for character phylogeny of a sequence position along the selected topology with the rates implied by the model,
d) multiplication of the probabilities gained from (b) for all sequence positions of the alignment

to obtain a likelihood value for the selected dendrogram.

e) The steps (c) and (d) are repeated for all alternative dendrograms in order to select the one with the best likelihood values.

The estimation of a probability value for the **character phylogeny** of a sequence position is the central problem of the method. For the given alignment and topology each sequence position can be considered separately assuming that characters evolve independently from each other. For a specific position in a given topology the character state distribution could be as follows:

The node sequences (ground patterns) ***X*** and ***Y*** are not known. However, only four character states can be taken into account, so that there are in total 16 alternative character phylogenies for this dendrogram when all possible ground patterns are considered:

X: A A A A G G G G C C C C T T T T
Y: A G C T A G C T A G C T A G C T

For each of these 16 alternatives the probability L_m can be calculated for the substitution of ground pattern states (characters in the nodes *X* and *Y*) to produce the characters of the terminal taxa. A biologically convenient assumption is that a substitution on one branch of the dendrogram is independent of the events on the other branches, including the directly preceding ones, and it is (usually) independent of the changes in other positions (**Markov**-model). The computation of the probability is model-dependent, the systematist has to select *a priori* a model that appears to be realistic for a specific gene or sequence region.

Using the "two-parameter-model" of Kimura for example (Kimura 1980; K2P-model; compare ch. 8.1, 14.1.3, Swofford et al. 1996), the probability $P_{(il)}$ for the occurrence of a character change of nucleotide i to nucleotide l is estimated with:

for $i=l$ (no change):

$$P_{il}(t) = \frac{1}{4} + \frac{1}{4} \cdot e^{-\lambda t} + \frac{1}{2} \cdot e^{-\lambda t \left(\frac{K+1}{2} \right)}$$

for transitions:

$$P_{il}(t) = \frac{1}{4} + \frac{1}{4} \cdot e^{-\lambda t} - \frac{1}{2} \cdot e^{-\lambda t \left(\frac{K+1}{2} \right)}$$

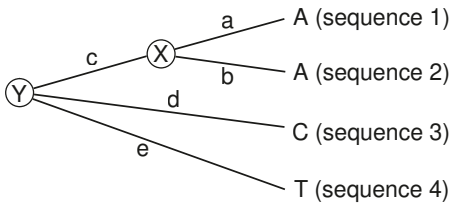


Fig. 203. Reconstruction of the evolution of a sequence position. ***X*** and ***Y*** are ground patterns of unknown character state, ***a***-***e*** are branch lengths which correspond to the number of substitution events.

for transversions:

$$P_{il}(t) = \frac{1}{4} - \frac{1}{4} \cdot e^{-\lambda t}$$

The axiomatic assumptions implied by this model are discussed in ch. 8.1 and 14.1.3. In these formulas λ is the average substitution rate of all bases and K is α/β (α is the substitution rate for transitions, β the rate for transversions).

The values for the model parameters $\lambda \cdot t$ and K have to be proposed or estimated with empirical data. For example, values for substitution rates could be adopted from other analyses that gave plausible results, or different rates are tested to find the ones that give the best likelihood values. Values for the ratio of α to β (transitions : transversions) can be gained by comparing terminal sequences (see ch. 8.2.6). The expression $\lambda \cdot t$ represents the most likely number of events per branch length and therefore also the substitution probability from one node to the next when the time between nodes is known.

Methodologically, if the time separating two nodes is not known from other sources, the parameters λ and t cannot be separated without assumptions on the existence of a universal molecular clock and a specific rate. The determination of the branch length $\lambda \cdot t$ can consist, for example, in testing different values for parameters with a computer program in order to select the values which maximize probabilities ("maximum likelihood") (see e.g., Tillier 1994, Lewis et al. 1996, Swofford et al. 1996). Remember that absolute rates are not needed for the reconstruction of the topology, because the relative ratio of branch lengths to each other together with information on closest neighbours already determines the topology.

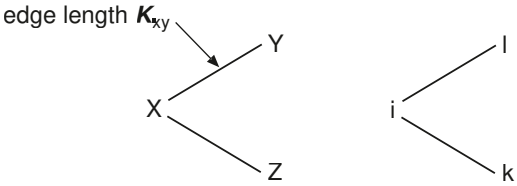


Fig. 204. Illustration of the evolution of sequences and of sequence positions for a section of a fictitious dendrogram. ***X***, ***Y*** and ***Z*** are complete sequences, ***i***, ***l*** and ***k*** individual nucleotides of sequence position ***j***. The branch length K_{XY} represents the relative length in the topology, not the length caused by the individual character. The probability that ***i*** is substituted by ***l*** is $P_{il}(K_{XY})$ when the nucleotide ***l*** is known.

The formulas listed above result from the K2P-model, their exact derivation is not explained here (but see models in ch. 14.3). Complex models allow the distinction of substitution probabilities for each specific type of substitution between nucleotides, summarized with a 4×4 matrix of substitution probabilities. Furthermore, it is possible to consider base frequencies and to assume that substitution rates are not constant but vary with the alignment position and with different branches of a dendrogram (ch. 8.1, 14.1).

For the estimation of the probability L_m of a given character phylogeny (Fig. 203) for each branch (***a*** to ***e***) of a dendrogram, the probability $P_{(il)}$ of a character change has to be considered by multiplying the ***P*** values which result from assuming specific nucleotides in ground pattern sequences (inner node characters). It has to be taken into account that there exist several alternatives for each inner node which result from all possible combinations of character states (16 alternatives for two inner nodes in Fig. 203).

Considering in Fig. 204 for sequence position ***j*** and dendrogram ***T*** the character states in an "ancestor" ***X*** and in its two daughter species ***Y*** and ***Z***, then the probability L_{Xi} that in node ***X*** nucleotide ***i*** occurs depends on whether character states ***l*** and ***k*** can evolve from ***i***.

The probability that nucleotide ***i*** is substituted by ***l*** on the branch length K_{XY} depends on the substitution process and the branch length, thus on P_{il} for $K_{XY,}$ as well as on the probability L_{Yl} that sequence ***Y*** actually shows the nucleotide ***l***. If ***Y*** is a terminal sequence, nucleotide ***l*** is known and for this ***l*** L_{Yl} has the value 1. Otherwise ***Y*** is an ancestral sequence for which in an earlier step the same calculation as for ***X*** has to be performed before a value L_{Yl} is available to calculate L_{Xi}. Therefore, the probability for the substitution $i \rightarrow l$ is $P_{il}(K_{XY})$ L_{Yl}, for $i \rightarrow k$ correspondingly $P_{ik}(K_{XZ})$ L_{Zk}. In this formulation K_{XY} is the branch length $\lambda_{XY}t_{XY}$, with a λ valid in the whole dendrogram, independently of the individual character phylogeny. For P_{il} the model-specific substitution probability is used, for example, P_{il} from the K2P-formulas mentioned above.

When there are more alternatives for *l* and *k* because *l* and *k* are unknown nucleotides of ground patterns, for each of all alternatives (= for each of four nucleotides ***l*** per position of the ground pattern) the probability has to be considered. Then the total probability for the substitution $i \rightarrow l$ is obtained with

$$\sum_{l} P_{il}(K_{XY})L_{Yl}$$

The value L_{Yl} has to be calculated in a preceding step analogous to the one for L_{Xi}, whereby statements on ground patterns are made along the topology in a descending (top-down) way. For the probability of having nucleotide *i* in node *X*, the probability for the two "descendant characters" *k* and *l* has to be considered too:

$$L_{Xi} = \sum_{l} P_{il}(K_{XY})L_{Yl} \cdot \sum_{k} P_{ik}(K_{XZ})L_{Zk}$$

This expression comprises the probability for the subtree of Fig. 204. With this principle, likelihood values can be calculated for the character phylogeny of a sequence position *j* in a given dendrogram *T*. The probability P_{il} is model-dependent, and the unknown branch lengths have to be estimated (see below) for the topology that is being considered. For the complete dendrogram *T* the likelihood L_j contributed by an alignment position *j* results from multiplication of all L_{Xi} values obtained for all nodes of the dendrogram.

This calculation has to be performed *for each position j* of the alignment and the likelihood values for all sequence positions have to be multiplied to obtain a likelihood for the topology. The probability for the given dendrogram *T* considering the complete alignment with *n* positions is obtained with

$$L_T = \prod_{j=1}^{n} L_j$$

As this value is very small, its logarithmic transformation is used for further topology comparisons:

$$\ln L = \sum_{j=1}^{n} L_j$$

For the unknown parameters (e.g., the branch length λt), which results from the multiplication of the time interval of a branch with the substitution rate), different values can be tested automatically with a computer program, calculating L for each variant. Those parameters for which the L value is maximal are selected.

This calculation is repeated for all possible dendrograms T which can be constructed from a dataset. The dendrogram with the highest value is chosen as the most probable one. The methods of dendrogram construction introduced in ch. 14.2.1 can be used for the construction and selection of topologies that are to be tested. As the L value has to be calculated for all possible topologies, ML-methods are very time consuming. A faster heuristic method of topology-search is possible with quartet-puzzling (see below).

Maximum-likelihood methods are very complex. The preceding explanations are only an introduction to the major principles of ML techniques. To learn further details it is recommended to study the original literature (e.g., Felsenstein 1981, Hasegawa et al. 1991, Mow 1994, Tillier 1994, Yang 1994).

Quartet-Puzzling

With this method, a maximum likelihood topology is calculated for each of the $\binom{n}{4}$ possible quartets which are combined from a set of n sequences (Strimmer & von Haeseler 1996). These partial trees are composed to a total topology ("puzzling"-step). This step is repeated many times and the final selected topology represents a consensus tree built from all optimal topologies. Furthermore, branch lengths and model parameters can be estimated for this topology.

For each quartet of sequences there are three possible dichotomous topologies (compare Fig. 60). For each of these three topologies a maximum likelihood value is calculated as explained above and the one with the maximal value is selected. Each selected quartet-topology for the taxa A, B, C, and D represents a split or a "neighbour relationship" {(A,B),(C,D)}. With biological data that usually contain homoplasies, through combination of optimal quartet-topologies one generally obtains a network (Fig. 196), wherefore approximation methods are used to get a dendrogram. By random selection of the first four sequences one gets the first quartet-topology. With the fifth randomly selected taxon E, it is tested which three of the first four taxa i, j, k, l show a neighbour relationship with E {(E,i),(j,k)}. Branches for which E does not appear as involved neighbour taxon are weighted. After testing all quartets which contain E and three taxa of the first topology, E is joined with the branch that has the lowest weight. If there are two equivalent solutions, one quartet is selected arbitrarily. After addition of all sequences the first topology is found, which is considered a provisional result.

Repeating these "puzzling" steps (for example 1000 times) also other alternatives are obtained. After several iterations usually a number of optimal, equivalent topologies are found, which are combined to a consensus tree. For each branch it can be stated how often it has been found in the independent puzzling steps. This value represents evidence for the support of a possible hypothesis of monophyly by the data in a way at first sight comparable to the bootstrap-value, however, puzzling does not imply a random selection of sequence positions but instead of different quartet combinations. Only clades which have a support clearly over 50 % should be discussed as serious candidates for monophyly hypotheses. To be added is the warning that all methods depict only the structure of the data, the interpretation of results has to be done by the biologist. The plesiomorphy trap is not detected with likelihood methods.

The method is significantly faster than the ML-technique, which requires complete topologies already at the beginning of the analysis (see above). If, however, the number of puzzling replicates is too small, the results will be unreliable.

Bayesian phylogeny inference is even faster than puzzling and seems to be a promising technique. The basics are explained in ch. 8.4.

14.7 Hadamard conjugation and Hendy-Penny spectra

The following problem is to be solved when an alignment is given: a topology should be found which matches the dataset best considering all possible splits which are present in the dataset, and with the help of models of sequence evolution which allow corrections for multiple substitutions.

In the following only the basic principles of the method are presented, which is based on the work of the mathematician J. Hadamard (1865-1963) and has been developed for the analysis of sequence evolution by Hendy and Penny (1993; see also Lento et al. 1995, Swofford et al. 1996). A Hadamard matrix is a simple matrix only consisting of the numbers 1 and –1, in which the pattern $\boldsymbol{H}$ is repeated. The basic element is the matrix of first order:

$$H^{(1)} = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix}$$

In a matrix of higher order this pattern is repeated, as illustrated for the second order pattern in the following example:

$$H^{(2)} = \begin{pmatrix} H^{(1)} & H^{(1)} \\ H^{(1)} & -H^{(1)} \end{pmatrix} = \begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{pmatrix}$$

With the help of the Hadamard matrix combinations of binary elements can be arranged per row and column. For spectral analysis in the sense of Hendy and Penny (1993), all possible splits (bipartitions in the set of species) have to be considered for a given number of species. When $\boldsymbol{T}$ is the number of taxa, maximally $\boldsymbol{m = 2^{T-1}}$ splits can occur (see also Fig. 151). For further calculations a Hadamard matrix is needed which has $\boldsymbol{m = 2^{T-1}}$ columns and rows. For four species there exist 8 splits when the split between the four species and the "rest of the world" is included, the corresponding Hadamard matrix therefore has 8 rows and columns (the "1" is omitted for clarity).

$$\begin{pmatrix} + & + & + & + & + & + & + & + \\ + & - & + & - & + & - & + & - \\ + & + & - & - & + & + & - & - \\ + & - & - & + & + & - & - & + \\ + & + & + & + & - & - & - & - \\ + & - & + & - & - & + & - & + \\ + & + & - & - & - & - & + & + \\ + & - & - & + & - & + & + & - \end{pmatrix}$$

It was shown by Hendy and Penny that this matrix is suited to describe all possible splits of a dataset, when for the naming of splits the following conventions are observed which allow the unequivocal identification of each split:

For each of the taxa t_1, t_2, t_3, t_4, … t_n a new index is introduced which is defined by 2^{i-1}, for example for t_1: 1; t_2: 2; t_3: 4; t_4: 8. For four taxa three dichotomous topologies can be constructed (compare Fig. 60), altogether with seven different splits and corresponding branches. The missing split with the branch $\boldsymbol{k_0}$ is the one to the "rest of the world". In order to name these branches $\boldsymbol{k}$ individually and unequivocally, the following convention is observed: the index of a branch results from the sum of the index numbers of the taxa present on that side of the split in which the highest species index does *not* occur.

Example:

Edge k_1 for the split {(1),(2,3,4)}:
The index 1 of the edge results from the species index "1".

Edge k_2 for the split {(2),(1,3,4)}:
The index 2 of the edge results from the species index "2".

Edge k_3 for the split {(1,2),(3,4)}:
the index 3 of the edge results from "1+2".

Edge k_4 for the split {(3),(1,2,4)}:
The index 4 of the edge results from the species index "4".

Edge k_5 for the split {(1,3),(2,4)}:
The index 5 of the edge results from "1+4".

Edge k_6 for the split {(1,4),(2,3)}:
The index 6 of the edge results from "2+4".

Edge k_7 for the split {(4),(1,2,3)}:
The index 7 of the edge results from "1+2+4".

Edge k_0 for the split {(),(1,2,3,4)}

For the topology of Fig. 205 the branch lengths can be described as a vector which in the order of edge indices has the following entries: –0.67; 0.3; 0.1; 0.05; 0.2; 0.0; 0.0; 0.02. The first entry for the edge k_0 is obtained as the negative value of the sum of all other branch lengths.

Another convention concerns the description of distances in a topology, which can be considered as the "path" between two terminal species or groups of species. The path from taxon 4 to taxon 2 (description of path {2,4}) consists of $k_7+k_3+k_2$ in the topology of Fig. 205. An unequivocal registration of all possible paths between terminal taxa is achieved with the following convention:

For each split describing an edge k_i, the group which does not contain the highest species index (see above) is considered, for example the group {2,3} for the branch k_6. The following rule is introduced: when the group contains an odd number of taxa then the last taxon (t_4 in this case) is inserted, if it is an even number the group remains as it is. With this convention one obtains for each edge a defined corresponding path description. The path description {1,2} is the path from taxon 1 to taxon 2. Attention: edge and path description are not the same! For four taxa the following edges and path descriptions result from this convention: k_1: {1,4}; k_2: {2,4}; k_3: {1,2}; k_4: {3,4}; k_5: {1,3}; k_6: {2,3}; k_7: {1,2,3,4}; k_0: {}. To understand which distances are included with this description you may colour these paths in Fig. 205!

This at first obscure convention is interesting because with the help of the Hadamard matrix and the description of all possible path lengths a relation between evolutionary branch lengths (estimated number of historical substitutions) and observed (countable) edge lengths can be constructed without having to refer to a specific topology. The product of edge lengths and matrix yields the lengths for the path descriptions corresponding to the branches.

For the example of Fig. 205 a Hadamard matrix with eight rows and columns has to be used (see above). Thereby the order of rows of the matrix is the same as the one for edge indices. The second row of the matrix thus corresponds to the edge k_1 and the corresponding path description is {1, 4}. This is equivalent to the sum $k_1+k_3+k_7$ in the

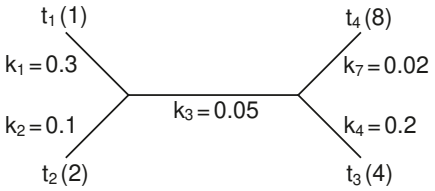


Fig. 205. Naming of taxa and edges for the Hadamard conjugation. The index ***t*** of the taxa is shown in brackets. The edge lengths in this fictive example add to a tree length of 0.67.

topology. This sum is obtained even when the topology is not known when edge lengths are given: the product of the second row of the Hadamard matrix with the vector of edge lengths results in:

$$-0.67 -0.3 +0.1 -0.05 +0.2 -0.0 +0.0 -0.02 = 0.74.$$

As the paths occur twice, the value calculated for the path length {1,4} has to be divided by two (0.74/2=0.37; see Fig. 205!). Recapitulating the whole procedure it becomes apparent that under observance of the conventions (and observance of the order of entries) the vector entries correspond to the length of specific edges, and that with the Hadamard conjugation defined path lengths between all terminal taxa in all possible topologies are obtained, without having to draw single topologies. All calculations are reversible and no information is lost.

Estimation of the expected spectrum of split-supporting positions

A histogram showing the support for single splits with the height of columns will be called a "spectrum" in the following. The support could be the estimated evolutionary distance between the two groups of a split, the observed number of supporting substitutions, or some other value (compare Fig. 170). We will first consider the Hendy-Penny method starting from a topology. Later the reverse way from the sequences to the topology can be explained. For a better understanding of the method it has to be explained first how the expected spectrum of supporting positions of all possible splits can be estimated for a *given* phylogenetic tree, when the *process of sequence evolution* is taken into account:

In a given gene tree the sequences of the terminal taxa and the reconstructed sequences of the ground patterns (= node sequences) are separated by a number of visible substitutions, which are interpreted as apomorphies. Since also analogous substitutions can occur, the considered sequences probably also contain some supporting positions for splits which are not compatible with the topology of the real phylogenetic tree (compare Figs. 55, 105, 150). In a pure dataset perfectly fitting to a topology such splits would not occur. In four terminal sequences up to eight different splits can occur (see Fig. 151) when trivial splits and the split between the species set and the "rest of the world" are taken into account. With T taxa there are maximally 2^{T-1} splits. Therefore, for a given dataset it can be expected that a large spectrum of supporting positions occurs, where some splits are formed by homologies, many however by analogies. To estimate the **expected spectrum** of split-supporting positions for a given gene tree the following steps are necessary:

- it is presupposed that a phylogenetic tree of the gene or sequences is given. The branch lengths of this tree represent the probabilities for character changes of a position along a branch. They serve to estimate the number of differences which the sequences probably have at the branch ends.
- The evolutionary branch lengths, which correspond to an estimation of the real number of substitution events, can be calculated from visible distances with a suitable model of sequence evolution as in distance methods (see ch. 8.2, 14.3).
- The last step in this consideration is the estimation of the expected spectrum of supporting positions for all splits, thus also for those splits which could originate by random, analogous character changes as a consequence of real substitution events. Starting from the given topology, the spectrum consists of the expected portions of sequence positions of an alignment which support individual splits (compare Fig. 170).

To estimate the probability that for a certain nontrivial split (compare Fig. 151) supporting positions occur, it is necessary to consider possible character states of ground patterns.

For the probability P that a nucleotide is substituted, the convention is that with $P=1$ a substitution always occurs along a branch and contrarily never with $P=0$. The probability that a character of a given ground pattern is substituted (change of the character states in Fig. 206: $0\rightarrow1$ or $1\rightarrow0$), corresponds to the respective branch length y (substitution rate multiplied by time). For the retention of the character state ($0\rightarrow0$ or $1\rightarrow1$) the probability is correspondingly *1-y*. Starting with one of the alternative constellations of ground pattern characters, the probability that the character states of the terminal taxa originate can be estimated by multiplication of the probability of character change of each branch. For the first pattern of node characters listed in Fig. 206 there is:

$$P = (1-y_a)\cdot(1-y_b)\cdot(1-y_c)\cdot y_d\cdot y_e$$

In order to estimate the probability $P_{(0,0/1,1)}$ that the character distribution of the terminal taxa of Fig. 206 is supported, the probabilities *for each possible combination of ground patterns* (in Fig. 206 there are four variations!) have to be added (compare maximum likelihood methods, ch. 14.6). As a specific split is clearly supported by two different character patterns of a data matrix or of an alignment, and indeed like in Fig. 206 by the pattern {(0,0),(1,1)}, but also by the pattern {(1,1),(0,0)}, the probability value $P_{(0,0/1,1)}$ just described has to be doubled. In the end this calculation has to be performed for each possible split of a dataset.

As the number of splits that have to be considered increases exponentially with increasing number of species, this calculation is not useful in practice. Hendy and Penny discovered that the Hadamard conjugation can be used to perform efficiently the model-dependent estimation of the spectrum of expected branch lengths just described (see Hendy & Penny 1993, Swofford et al. 1996).

First a vector p is defined which contains m elements, whereby each element is the branch length of a split (in the sense of Fig. 206). Here m is the number of all possible splits ($m=2^{T-1}$) for T taxa. Which splits are considered at first can be selected at random, but the order then has to be maintained for the remaining steps. The vector p has the following form:

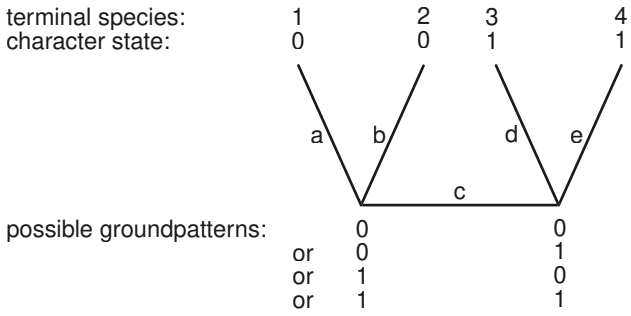


Fig. 206. Possible ground patterns for a topology with supporting binary character states for the illustrated split {(1,2),(3,4)}. The letters ***a-e*** name the edges and their length p_i.

$$p = \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \cdots \\ p_{m-1} \end{pmatrix}$$

The edge p_0 corresponds to the split between the considered group of species and the "rest of the world", for which no data are present, and it has the value 0. There remain m–1 splits. How the branches are named has already been explained above.

Example: for four species there are eight possible splits (Fig. 151). When the dataset supports unequivocally only one dichotomous topology, there are positive edge lengths only for five edges. Those splits which do not occur in the dataset get the edge length 0 (Fig. 207).

As in distance methods, these edge lengths p_i can be considered to be **visible distances** from which **additive "evolutionary distances"** q_i can be calculated with the help of models of sequence evolution, which enable corrections for multiple substitutions and expected chance similarities. The Jukes-Cantor correction (Poisson-correction) for binary characters for example runs as follows: $q_i = -0.5 \ln(1-2p_i)$ (compare ch. 14.1.1). Replacing p_i by q_i, in each case in the vector p, a *spectrum* γ *of "evolutionary branch lengths"* is obtained (γ vector, topology or tree spectrum). For q_0 the sum of the other q_i-values is entered with a negative sign. Attention: it is no problem to recover again p_i from the q_i entries by backward calculation. The method is reversible.

For ideal data the vector γ has as many positive entries as the corresponding topology has branches. The entry "0" means that for a possible split no branch is present in the topology. In place of such a zero-split, a positive value will be found in real data when analogies occur.

From this γ-vector the *spectrum* ***S*** *of expected branch lengths* of all partitions of the species of a dataset can be obtained, which is the spectrum of supporting positions that can occur in an alignment when sequence evolution proceeded according to the selected model

$$S = H^{-1} \cdot \exp[H \cdot \gamma]$$

Here $H \cdot \gamma$ stands for the multiplication of the Hadamard matrix, which has $m = 2^{T-1}$ columns and rows, with the vector γ (vector of evolutionary path lengths). The multiplication $H \cdot \gamma$ yields a further vector rho (ρ), which shows the path lengths of all possible paths between terminal taxa of the dataset (see above: explanation of the Hadamard conjugation). The distances in the rho-vector correspond to evolutionary distances, be-

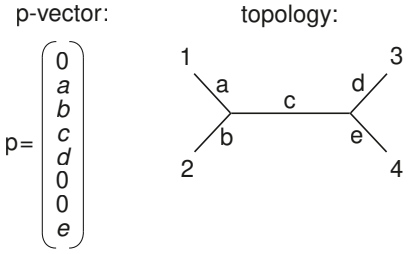


Fig. 207. Convention to record branch lengths of a topology with a vector: the splits {(1,3),(2,4)} and {(2,3),(1,4)} are not represented in the topology.

cause they are obtained from the vector γ. With the exponential function ***exp [H·γ]*** the correction for multiple substitutions is reversed and "visible" distances are obtained again. The expression ***exp[ρ]*** consists of the elements e^i (for p_0 to p_{m-1}, p_i = visible distance) and corresponds to the vector ***r***. Remember: with these vectors (r and rho) not only are pairwise distances or distances on a tree (in contrast to *p* and γ) described, but all paths between terminal taxa, which is why they are called "generalized distances". Through reversal of the Hadamard conjugation (formula above) the spectrum *S* of supporting positions is obtained ("sequence spectrum").

From the sequence spectrum neither the alignment nor the positions which support a group can be reconstructed. The sequence spectrum obtained with the Hadamard conjugation represents the support for single splits as a portion of the whole data matrix, and the *sum of all portions is 1*. The conserved, invariable positions support the first "split", which unites all species of the dataset in one group. Furthermore also all trivial splits are present which are supported by autapomorphies of individual species. For the formation of groups only those splits which contain groups of at least two terminal species (or sequences) are interesting (compare Fig. 151).

A special feature of this method is that all calculations can be easily reversed. To review all steps:

Topology ↔ visible ***p***-distances of the branches in the *p*-vector ↔ (transformation with model of evolution) ↔ evolutionary ***q***-distances in the γ-vector (tree spectrum) ↔ (Hadamard-conjugation) ↔ ***rho***-vector (with generalized evolutionary distances) ↔ (transformation with model of evolution) ↔ ***r***-vector (with visible or uncorrected generalized distances) ↔ (Hadamard-conjugation) ↔ ***S***-Spectrum (sequence spectrum; portion of supporting positions for splits).

Attention: the calculation of the spectrum ***S*** by modification of the γ-vector with the variation of the above described Hadamard conjugation is only valid under the condition that sequence evolution corresponds to the Jukes-Cantor model and that binary characters are used! For four character states and the K3ST-model (compare Fig. 159) variations of the above described method were suggested (see Hendy et al. 1994, Swofford et al. 1996). In each case all assumptions implied by the models used have to be realistic for the selected sequences.

Evaluation of a real spectrum of supporting positions

It has been described in the preceding paragraph how the *expected spectrum* can be estimated assuming that a tree is known. In practice, however, the reverse case is relevant: an alignment contains a number of splits supported by sequence positions and we want to find the best fitting tree. The Hadamard conjugation can also be used for this purpose. From an *observed spectrum* of supporting positions ***S***' (supporting positions of all bipartitions in an alignment) the estimated spectrum of branch lengths of a phylogenetic tree is obtained with

$$\gamma' = H^{-1} \cdot \ln (H \cdot S')$$

This γ'-vector (reconstructed tree spectrum) is an estimation of the number of real substitution events along branches of a tree. If the spectrum ***S'*** is an *exact* sample of the support for the splits and the Jukes-Cantor model represents the real process of sequence evolution, then one could also obtain from the vector γ the exact number of visible sequence differences $\boldsymbol{p}_i$ reversing the steps explained above. However, in each real dataset some errors occur with high probability: one should not expect that the spectrum ***S***' corresponds exactly to the predicted spectrum ***S*** of the preceding paragraph, because real data usually only represent an imperfect sample and real rates of sequence evolution possibly will not have the constant and uniform values which are required by the selected model of evolution.

Often many splits of a real dataset are only supported by character states that are identical due to analogous substitutions and they form "false signals", while others are "real signals" supported by real apomorphies. Since homology signal is not random, it should accumulate faster in large datasets than noise (Fig. 154). Assuming that in large and informative datasets the signals consisting of apomorphies are more distinct than chance similarities ("the background noise"), a spectrum of positions which support individual splits can be gained to study support and contra-

dictions. The spectrum can be used to estimate which edge lengths fit clearly to a dendrogram. In practice the following steps are performed:

First the frequency of support for individual splits is counted, the sum of all frequencies is 1. This means that the bipartitions contained in each alignment position must be recorded to get the number of supporting positions for each split represented in the alignment. These frequencies correspond to the spectrum $\boldsymbol{S}'$. Starting with it, the spectrum γ' can be obtained by reversing the calculation of the preceding chapter. Spectrum γ' corresponds to an estimation of the *expected* substitutions for all branches, it is proportional to the evolutionary branch lengths in a topology. From the spectrum γ' the vector $\boldsymbol{p}'$ can be derived, which in the ideal case corresponds to the observed (uncorrected) branch lengths. When informative data contain sampling errors or analogies that could not be cleared with the selected model, some splits will be found that are incompatible with the real phylogenetic tree. Ideally, their γ'-value should be close to 0. However, when high γ'-values for incompatible splits are found, either the model of sequence evolution is not suitable or the dataset is so noisy that the correct phylogenetic signal is not detectable.

A spectrum of γ'-values is represented graphically in form of a bar graph ("*Lento*-diagram", Fig. 170). For each split $\boldsymbol{X}$ not only the support can be visualized with the height of the bar, but also the "**conflict**", which is shown below the support bar. Conflict is a measure for the support for alternative groupings that are represented in the data, and which are *incompatible* with the split $\boldsymbol{X}$ (compare Figs. 151, 170). The support for these alternatives are added and drawn on the X-axis below the corresponding split column. Because each split can be incompatible with many other ones, conflict values can be much higher than support. Therefore Lento et al. (1995) proposed to normalize the frequency of conflicting splits by dividing the sum of all support values by the sum of all conflict values. Multiplication of each conflict value with this ratio gives a reasonable scale to draw an illustrative spectrum.

14.8 Relative rate test

All distances between terminal taxa are ultrametric if substitution rates in all lineages of the corresponding phylogeny are identical. This condition is probably very rare in nature. Presence of differences in substitution rates are the motivation to avoid simple clustering methods when trees are constructed from a matrix with pairwise distances (ch. 8.2.2, 14.3.7). Furthermore, high substitution rates of individual taxa are also a risk when other methods of tree inference are used (see "long branch problems": ch. 6.3.2, 8.2.3). The relative rate test can be performed to find out if rates deviate much from each other:

- Choose a reference species which definitely does not belong to the analysed ingroup.
- Determine pairwise distances between the species of the ingroup and the reference species.
- Rely on the assumption that when all substitution rates are similar, distances to the reference species should also be similar.

The implications of these statements are illustrated in Fig. 208.

In this test the autapomorphies of the *reference species* have no obvious influence on the comparison of distances, they are contained in each distance value and therefore subtracted when distance differences are compared. However, the synapomorphies and analogies shared by two terminal taxa (new states shared by species A+B in Fig. 208) and autapomorphies of terminal species have an effect. Also, it should not be ignored that analogies, back mutations, and multiple substitutions influence these values by reduction of observed distances. Some autapomorphies of the reference species may be homoplasious and they will have a distorting effect if base composition differs in lineages.

To get test results one can proceed as follows:

Define S_{13} (and accordingly S_{23}) as the average number of observed substitutions (when dealing with closely related species) or as the evolution-

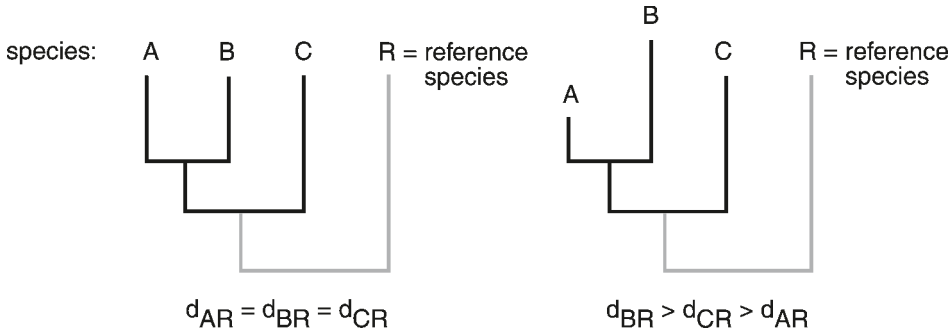


Fig. 208. Relative rate test: the branch lengths represent the number of substitutions which occurred in the time between the root or inner nodes and terminal taxa. Substitution rates are unequal in the example represented on the right side.

ary distance between species 1 (and correspondingly species 2) and the outgroup 3. The difference S_{13} minus S_{23} is positive when species 1 evolved faster than species 2, it is negative when species 1 evolved slower than species 2. The difference is considered to be significant when it is larger than twice the standard deviation estimated from the available samples (distances between single sequences and the selected outgroup). The variance of the rate test is described with the Z-score:

$$Z = \frac{S_{13} - S_{23}}{\sqrt{\mathrm{var}(S_{13} - S_{23})}}$$

The variance var $(S_{13} - S_{23})$ can be estimated analytically (Wu & Li 1985) or using the bootstrap resampling technique (Nei & Kumar 2000). Species pairs with highest absolute Z-values are those that deviate most from the average rate of the data set.

A good method to visualize taxon-specific rate differences is to plot Z-values on the Y-axis and the evolutionary distance values for the corresponding species pair on the X-axis. One can perform this analysis separately for different substitution types (transitions, third codon positions etc.; examples for applications: Wu & Li 1985, Lyrholm et al. 1990, Friedrich & Tautz 1997).

The assumption that comparable distances found in the relative rate test can be attributed to similar rates does not always prove true. A high number of multiple substitutions in lineages of the tree cannot be detected with the rate test. When the alignment is "saturated" in variable positions the true rate differences will not be discernible (Philippe & Laurent 1998). A consequence of false confidence in the relative rate test can be that taxa are erroneously identified as sister groups on the basis of analogies (a result of hidden long branches) and remaining taxa are joined by symplesiomorphies, or that the age of a taxon is not inferred correctly.

Further methods that have been suggested for the identification of differing substitution rates include the two-cluster-test, in which the average substitution rate is compared in two sister taxa, and the branch-length test in which the length of a stem lineage from the root to a terminal taxon is compared to the average length of all terminal taxa of a topology (Takezaki et al. 1995).

14.9 Evaluation of the information content of datasets using permutations

Permutation tests can indicate the structure of the data is not accidental and thus probably contains phylogenetic information. For cladistic tests the permutation is obtained by combining the terminal taxa of a dataset to construct random topologies and then tree length distribution is determined, or by scattering repeatedly the character states at random over the taxa in a matrix to produce artificial data matrices in which the number of characters and of their states is identical to the real data matrix.

PTP-Test (permutation tail probability test)

Applying the MP-method, randomized data can certainly yield an apparently unequivocal result when only one single shortest tree is found. This can even be markedly shorter than the next suboptimal tree. The difference to phylogenetically structured datasets is the frequency of lengths of different trees chosen at random from the tree space. The frequency should show a symmetrical distribution when randomized data are used,

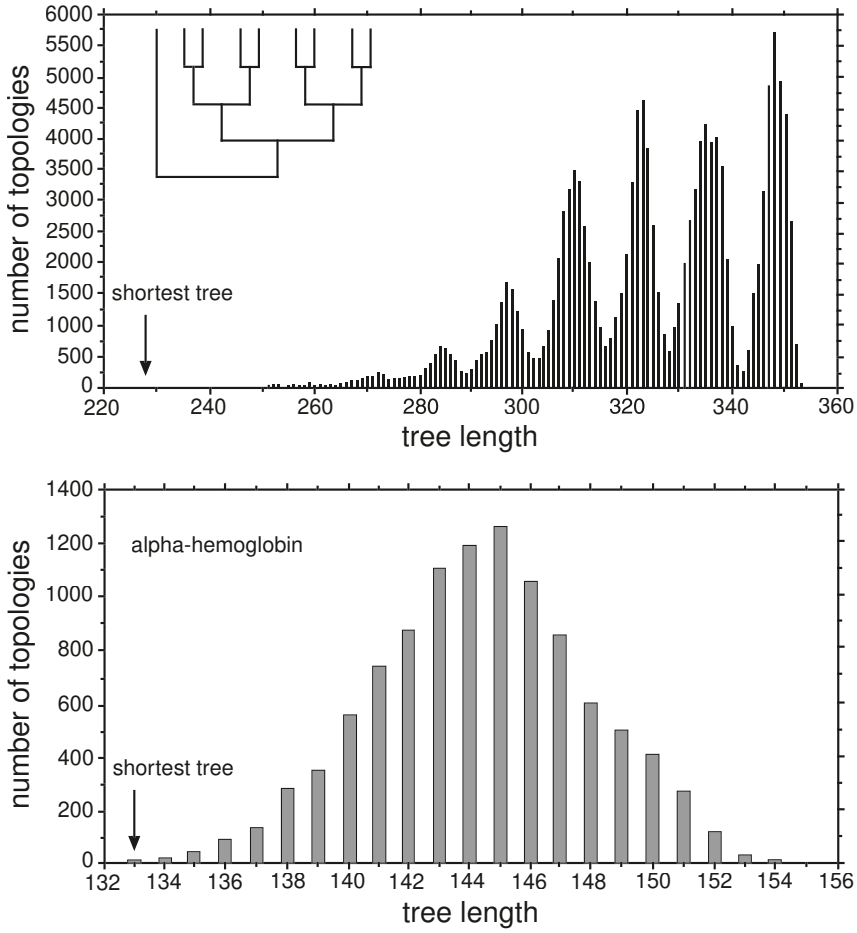


Fig. 209. Comparison of the tree length distribution of noisy character sets or of random data with phylogenetically structured datasets. Top: phylogeny of a laboratory culture of a bacteriophage. One obtains a non-random distribution of tree lengths from this dataset. Bottom: alpha-hemoglobin data with noisy, unstructured patterns of shared character states (after Hillis & Huelsenbeck 1992).

number of taxa:	5	5	6	6	7	7	8	8	9	9	10	10	15	15	20	20
p =	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1
number of characters																
10	–1.12	–1.30	–0.75	–1.02	–0.63	–0.84	–0.37	–0.67	–0.48	–0.56	–0.44	–0.59	–0.28	–0.37	–0.18	–0.24
50	–0.88	–1.08	–0.67	–0.88	–0.39	–0.63	–0.37	–0.49	–0.31	–0.44	–0.35	–0.39	–0.16	–0.20	–0.10	–0.11
100	–0.77	–1.08	–0.59	–0.68	–0.37	–0.46	–0.37	–0.43	–0.33	–0.43	–0.26	–0.31	–0.15	–0.19	–0.09	–0.10
250	–0.94	–1.20	–0.74	–1.12	–0.37	–0.49	–0.33	–0.44	–0.29	–0.44	–0.22	–0.35	–0.15	–0.20	–0.08	–0.09
500	–0.60	–0.84	–0.53	–0.63	–0.35	–0.46	–0.31	–0.47	–0.29	–0.47	–0.20	–0.27	–0.10	–0.15	–0.08	–0.08

Fig. 210. $\mathbf{g_1}$-values for binary characters (characters with only 2 character states, after Hillis & Huelsenbeck 1992; $\mathbf{g_1}$ values for four character states are found in the same publication).

whereas with non-random data more shorter than longer trees are obtained in comparison with the average tree length (Fig. 209). The reason for this is the accumulation of homologous character state changes on the few branches of the true topology, or in other words, the covariance of different homologies (Faith & Cranston 1991), while other topologies more different to the true tree would show a larger number of homoplasies (additional analogous character state changes). If there is no covariance, the tree length frequency would be symmetrical in the ideal case. The null hypothesis is that the shortest tree is contained in the symmetric distribution of random trees.

In order to describe the structure of data, Hillis & Huelsenbeck (1992) suggest the use of **"$\mathbf{g_1}$ statistics"** (Sokal & Rohlf 1981) which evaluates the asymmetry (skewedness) of the tree length distribution. Define ***n*** as the number of trees; ***s*** is the standard deviation for their tree lengths. For each of the tree lengths T_i the difference to the average of the tree length of all topologies is calculated. The symmetry of length frequencies is described with

$$g_1 = \frac{\sum_{i=1}^{T} \left(T_i - \overline{T}\right)^3}{ns^3}$$

For length frequencies with exactly symmetrical distribution $\mathbf{g_1}$ is 0, because in the numerator the number of trees which are shorter than the average is the same as the number of longer trees. When frequencies are shifted towards values larger than the average, $\mathbf{g_1}$ is <0, with lower values $\mathbf{g_1}$ is >0. With $\mathbf{g_1}<0$ it can be assumed that there are not many alternatives to the optimal solution (the shortest dendrogram), and that characters are correlated in a non-random way. The confidence value ***p*** listed in the table (Fig. 210) states that tree lengths deviate from a random distribution with 95 % or 99 % probability when g_1 is more negative than the value given in the table.

Attention:
- probability statements derived from a PTP-test do not concern the probability that the shortest MP tree depicts the true phylogeny, but only that the data contain non-random patterns.
- Statements about the quality of individual characters are not possible.
- The support or probability of monophyly of individual groups of terminal taxa cannot be determined.
- Using DNA sequences, the number of character changes which influence tree lengths does not depend on sequence length but on the number of variable positions.
- As some sequence regions can be more informative than others, it is worthwhile to perform the test separately for individual regions. However, to analyse the phylogenetic information content a more detailed insight is obtained with spectra of supporting positions which can be studied for single sequence areas and visualize the support for groups of species (ch. 6.5).

T-PTP-test (topology-dependent PTP-test)

The test described above does not allow statements about how well the available characters support individual clades. However, one can also perform the PTP-test for individual groups, by choosing those shortest topologies which contain a specific monophylum (or constructing a constraint tree that contains this group) and comparing the tree length with other random topologies in which the same taxa do not form a monophylum (Faith 1991). In analogy to simple PTP-tests

it is assumed that the length difference between the shortest tree containing the monophylum and the shortest tree without the monophylum can probably not be obtained by chance when the real monophylum is supported by several real homologies.

The significance of this difference is determined by comparing the difference in length obtained with the original data set with tree length difference distribution in a randomized data set. The randomized data are obtained by keeping the outgroup character states constant, while ingroup character states are permuted among taxa of the ingroup.

For example, Halanych (1995) got with an 18S-rDNA alignment a cluster composed of sequences of Pterobranchia and Enteropneusta. The single most parsimonious tree with this clade was 205 steps long, while the shortest tree without this clade had 209 steps, the length difference is 4. The calculation was repeated with randomized data and for each artificial alignment the length differences between the shortest trees with and shortest trees without this clade was recorded. A plot of these length differences (frequencies of tree length difference on the y axis, length difference values on the x axis) showed that the value four is rare in the randomized data sets, it falls within the 1 % tail of the distribution. Therefore, the author concluded that the data are significantly more structured than random data ($p \leq 0.01$).

Whereas in the randomization test described above the names of the taxa can be ignored in order to obtain the tree length frequencies, the analysis of the support for single potential monophyla requires the identification of terminal taxa.

Remember that this test will not tell you if the characters supporting monophyly are real homologies or, for example, the result of a bias in nucleotide frequencies.

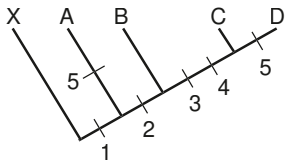
14.10 F-ratio

The F-ratio can be used for the comparison of dendrograms with a data matrix (Farris 1972, Brooks et al. 1986). The ratio is a measure for the number of homoplasies and can be used as criterion to choose between alternative dendrograms. A polarization (rooting) of the dendrograms is not required (Fig. 211).

data matrix

	characters				
species	1	2	3	4	5
X	0	0	0	0	0
A	1	0	0	0	1
B	1	1	0	0	0
C	1	1	1	1	0
D	1	1	1	1	1

selected dendrogram



phenetic distances

	X	A	B	C
X				
A	2			
B	2	2		
C	4	4	2	
D	5	3	3	1

branch lengths

	X	A	B	C
X				
A	2			
B	2	2		
C	4	4	2	
D	5	**5**	3	1

Fig. 211. Example for the calculation of the F-ratio (after Brooks et al. 1996). The dendrogram shows on which branches the numbered characters change their state. The phenetic distance indicates how many character states are different between pairs of species. The branch length is the sum of all character changes on the path connecting terminal species.

In this example the branch length (also *patristic distance*) between the species pair A-D is larger than the phenetic distance. The difference δ between the two matrices (sum of distances in matrix 1 minus sum of distances in matrix 2) corresponds to the number of homoplasies in the given topology and amounts to 2 steps. The sum S of the phenetic distances of the matrix is in this example 28, the **f-ratio is** 7.14:

$$F = (\delta / S) \cdot 100$$

If one of the otherwise identical dendrograms has more synapomorphies, the f-ratio is lower for this dendrogram. In the same way the value decreases with addition of autapomorphies (!) of terminal taxa. Reversals and analogies increase the value. The lowest f-ratio is not necessarily present in the shortest dendrograms.

The f-ratio can be calculated for discrete characters as in the example above, but also for continuous characters (e.g., immunological distances), from which branch lengths were calculated.

14.11 PAM-matrix

Substitution patterns of amino acid sequences are more complex than for nucleotides and are mostly empirical, not nested (for nested models of nucleotide substitutions see Fig. 160). Tables of amino acids serve in the estimation of the similarity of amino acid sequences, which cannot be obtained simply by comparing the number of identical amino acids. The similarity values can consider the chemical similarity of amino acids, the molecular structure of the protein, the substitution probability, and also the phylogenetic relationships of organisms containing the proteins.

The PAM-matrix reproduced here was the first that became popular. It is based on data acquired prior to 1978 and is therefore not up-to-date. It requires the assumption that all positions are equally variable. Furthermore it was derived from relatively small proteins of closely related species, wherefore they are not suited for many phylogenetic studies of distantly related species. The interested reader can find suggestions that avoid the disadvantages of the PAM-matrix in the recent literature (or web sites; compare Gonnet et al. 1992; BLOSUM matrix (block substitution

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	2	0	–2	0	0	–4	1	–1	–1	–1	–2	–1	0	1	0	–2	1	1	0	–6	–3	0
B	0	2	–4	3	2	–5	0	1	–2	1	–3	–2	2	–1	1	–1	0	0	–2	–5	–3	2
C	–2	–4	12	–5	–5	–4	–3	–3	–2	–5	–6	–5	–4	–3	–5	–4	0	–2	–2	–8	0	–5
D	0	3	–5	4	3	–6	1	1	–2	0	–4	–3	2	–1	2	–1	0	0	–2	–7	–4	3
E	0	2	–5	3	4	–5	0	1	–2	0	–3	–2	1	–1	2	–1	0	0	–2	–7	–4	3
F	–4	–5	–4	–6	–5	9	–5	–2	1	–5	2	0	–4	–5	–5	–4	–3	–3	–1	0	7	–5
G	1	0	–3	1	0	–5	5	–2	–3	–2	–4	–3	0	–1	–1	–3	1	0	–1	–7	–5	–1
H	–1	1	–3	1	1	–2	–2	6	–2	0	–2	–2	2	0	3	2	–1	–1	–2	–3	0	2
I	–1	–2	–2	–2	–2	1	–3	–2	5	–2	2	2	–2	–2	–2	–2	–1	0	4	–5	–1	–2
K	–1	1	–5	0	0	–5	–2	0	–2	5	–3	0	1	–1	1	3	0	0	–2	–3	–4	0
L	–2	–3	–6	–4	–3	2	–4	–2	2	–3	6	4	–3	–3	–2	–3	–3	–2	2	–2	–1	–3
M	–1	–2	–5	–3	–2	0	–3	–2	2	0	4	6	–2	–2	–1	0	–2	–1	2	–4	–2	–2
N	0	2	–4	2	1	–4	0	2	–2	1	–3	–2	2	–1	1	0	1	0	–2	–4	–2	1
P	1	–1	–3	–1	–1	–5	–1	0	–2	–1	–3	–2	–1	6	0	0	1	0	–1	–6	–5	0
Q	0	1	–5	2	2	–5	–1	3	–2	1	–2	–1	1	0	4	1	–1	–1	–2	–5	–4	3
R	–2	–1	–4	–1	–1	–4	–3	2	–2	3	–3	0	0	0	1	6	0	–1	–2	2	–4	0
S	1	0	0	0	0	–3	1	–1	–1	0	–3	–2	1	1	–1	0	2	1	–1	–2	–3	0
T	1	0	–2	0	0	–3	0	–1	0	0	–2	–1	0	0	–1	–1	1	3	0	–5	–3	–1
V	0	–2	–2	–2	–2	–1	–1	–2	4	–2	2	2	–2	–1	–2	–2	–1	0	4	–6	–2	–2
W	–6	–5	–8	–7	–7	0	–7	–3	–5	–3	–2	–4	–4	–6	–5	2	–2	–5	–6	17	0	–6
Y	–3	–3	0	–4	–4	7	–5	0	–1	–4	–1	–2	–2	–5	–4	–4	–3	–3	–2	0	10	–4
Z	0	2	–5	3	3	–5	–1	2	–2	0	–3	–2	1	0	3	0	0	–1	–2	–6	–4	3

Fig. 212. A PAM matrix used to determine the relative probabilities for substitutions of amino acids.

matrix): Henikoff & Henikoff 1992). It is possible, for example, to homologize positions on the basis of the three dimensional protein structure and to estimate the expected frequency for a change of configuration caused by amino acid substitutions. Selection would result in a low probability for such substitutions.

This matrix contains empirically found frequencies of mutations of amino acid sequences (Dayhoff et al. 1978, compare ch. 5.2.2.9), it indicates probabilities for the replacement of one specific amino acid by another one (without multiple substitutions). The symbols in the outer left column represent the original amino acid, the top row contains the replacements. To compile these data, 71 sequence families, of which the phylogeny had been reconstructed, were considered. The PAM-units ("point accepted mutations per 100 residues per 10^8 modelled evolutionary years") have a negative value when a replacement occurs empirically less often than expected in a random combination of amino acids. Positive numbers characterize pairs of amino acids which were observed more frequently than expected by chance alone. When the same amino acid is present in the same position in sequence pairs, the value indicates how probable the conservation of a plesiomorphic state is.

If amino acid substitution models rely on this matrix the following assumptions are implied: the replacement probability depends only on the value found in this table, substitution probabilities are similar in all regions of a sequence, and the sequences of an alignment have the same average amino acid composition. Note that with increasing evolutionary distance the reliability of the matrix decreases.

14.12 Optimization alignment

Optimization alignment allows a most parsimonious dendrogram to be obtained from raw data (unaligned sequences) with a method that guarantees the application of the same criteria to different datasets (Wheeler 1996, 2002). The main steps of the procedure are:

- take any topology that can be constructed from the raw data and terminal taxa, and select a step matrix for the substitution probabilities (= weights for character state changes);
- construct node sequences descending from the terminal taxa along the topology and adjust the alignment in each node (***down-pass***),
- select the node sequence(s)/alignment(s) that give the shortest branch lengths, using a *step matrix* (cost matrix, Fig. 141) that represents a model of sequence evolution,
- after arriving at the root, select the final node sequences in an ***up-pass*** through the topology (from the root to the terminal taxa),
- determine the ***tree length***.
- The resulting alignment is topology dependent, wherefore other topologies have to be searched (in the same way as in maximum parsimony analyses, in most cases using heuristic methods) to find the most parsimonious solution (which is an alignment and a corresponding tree).

The weight matrix (Fig. 213) has a strong influence on the result, as the optimal topology can

step matrix:

0	1	2	1	2
1	0	1	2	2
2	1	0	1	2
1	2	1	0	2
2	2	2	2	0

convention:

	A	**C**	**G**	**T**	**()**
A	0	1	2	1	2
C	1	0	1	2	2
G	2	1	0	1	2
T	1	2	1	0	2
()	2	2	2	2	0

Fig. 213. Example for a step matrix used for optimization alignment. Gaps are coded as fifth character states. At left: a matrix that is read by the program POY. At right: the corresponding convention. In this particular case the transitions (A-G, C-T) and indels (weight 2) increase tree length more than transversions (A-C, A-T, C-G, G-T).

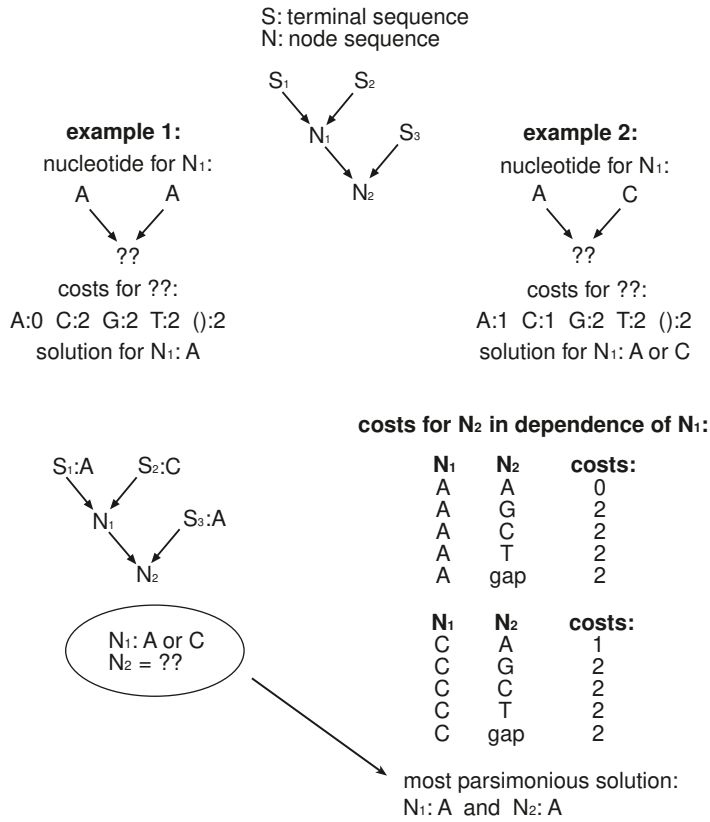


Fig. 214. Calculation of costs for internal node nucleotides assuming that each character state change has the weight '1' (see Wheeler 2002).

vary depending on the weights selected for character transformations.

Given a topology, the sequences and a step matrix, preliminary states are assigned to internal nodes during a *down-pass* (from the leaves to the root of the tree). Those states of internal nodes are selected that minimize tree length (Fig. 214).

Note that gaps are introduced whenever these minimize the cost (which is also the branch length in MP analyses). Working down the tree the exact solution would be to consider at all times all possible alternative states for each position in each inner node. Since this is computationally too time consuming, in a simplified approach costs are optimized by only considering neighbouring nodes (as in Fig. 214). To code for alternative nucleotides the IUPAC code can be used (Fig. 215), so that for each inner node 31 different possibilities exist.

Since the terminal sequences can vary in length, it can be unknown which of the positions of two sequences should be compared to infer a node character state. To find the optimal positional homology, the optimization alignment algorithm will consider different combinations with and without gaps to select the positions that minimize the costs. A simple example is explained in Fig. 216.

For the calculation of the cost of a character transformation, the path yielding the lowest cost is selected (Fig. 217).

The down-pass ends when states at the root of the tree are determined. At this stage many nodes may still contain suboptimal solutions. The final states are determined with an up-pass.

The up-pass starts at the root and works up to the leaves of the tree. Again, only neighbouring nodes

nucleic acids:		amino acids:		
code	description	code (one letter)	code (three letters)	description
A	Adenine	A	Ala	Alanine
C	Cytosine	R	Arg	Arginine
G	Guanine	N	Asn	Asparagine
T	Thymine	D	Asp	Aspartic acid
U	Uracil	C	Cys	Cysteine
R	Purine (A or G)	Q	Gln	Glutamine
Y	Pyrimidine (C, T, or U)	E	Glu	Glutamic acid
M	C or A	G	Gly	Glycine
K	T, U, or G	H	His	Histidine
W	T, U, or A	I	Ile	Isoleucine
S	C or G	L	Leu	Leucine
B	C, T, U, or G (not A)	K	Lys	Lysine
D	A, T, U, or G (not C)	M	Met	Methionine
H	A, T, U, or C (not G)	F	Phe	Phenylalanine
V	A, C, or G (not T, not U)	P	Pro	Proline
N	Any base (A, C, G, T, or U)	S	Ser	Serine
		T	Thr	Threonine
		W	Trp	Tryptophan
		Y	Tyr	Tyrosine
		V	Val	Valine
		B	Asx	Aspartic acid or Asparagine
		Z	Glx	Glutamine or Glutamic acid
		X	Xaa	Any amino acid

Fig. 215. The IUPAC code for nucleotides and amino acids. In addition to the 15 symbols for *nucleotides* another 15 are the combination with a gap (e.g., (M)), and adding the gap symbol we get a total of 31 symbols.

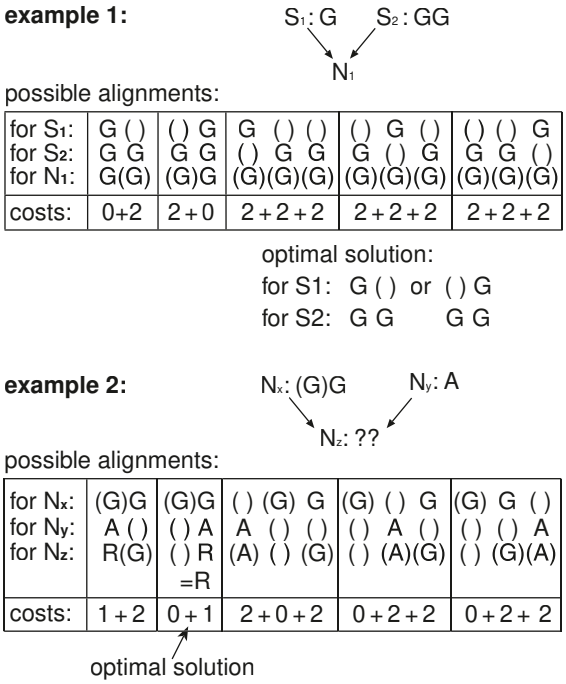


Fig. 216. Optimization of an alignment for two neighbouring terminal sequences and the corresponding inner node, assuming that each substitution has the cost '1' and an indel costs '2'.

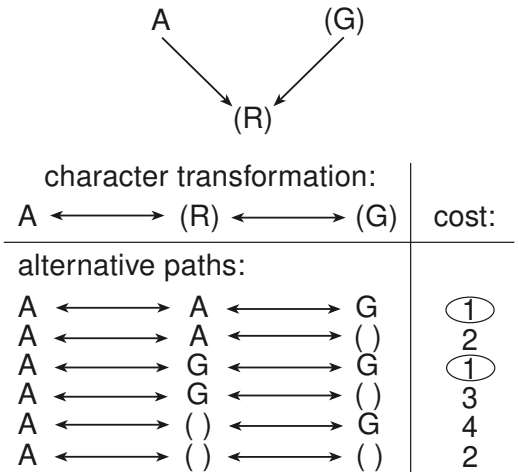
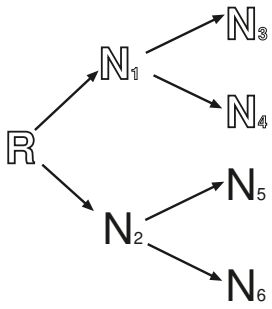


Fig. 217. Estimation of the cost for a character transformation when there are several alternatives for character states. Note that (R) means 'A, G, or ()'. There are two optimal solutions, the total cost is '1'.



determination of character state for $\mathbb{N}_1$:

states from down-pass	costs for $\mathbb{N}_1$ = A:	costs for $\mathbb{N}_1$ = G:
ℝ = A	0	1
$\mathbb{N}_1$ = A or G	-	-
$\mathbb{N}_3$ = G or (G)	G: 1 ():2	G: 0 ():2
$\mathbb{N}_4$ = A	0	1
smallest sum:	1	2

Fig. 218. Determination of node character states during the up-pass, assuming that substitutions have the cost '1' and indels the cost '2'. The most parsimonious state for node N_1 is A.

are considered to minimize computation costs (locally low cost reconstruction). An example is shown in Fig. 218.

An important question is whether or not optimization alignment leads to a circular argument. This is certainly not the case, because different trees and alignments are tested and the result is not contained within the first assumptions of the method. However, alignment and trees depend on the selected weight matrix and the implied assumption that the matrix is valid for all parts of the tree.

In summary, some characteristic features of optimization alignment are:

- a topology is constructed from unaligned sequences using the principle of parsimony,
- a large number of alignments is considered to select the most parsimonious solution,
- the resulting topology depends on the selected weight matrix,
- the user does not interfere to eliminate ambiguous alignment regions,
- with large datasets a heuristic tree search is necessary,
- weighting of single positions in dependence of the variability of sequence regions is not possible,
- gaps are treated as fifth character state,
- computation time is much larger than for MP analyses with pre-aligned sequences.

Some of these features may change with future modifications of the algorithms.

15. Available computer programs, web sites

A recommendable "cookbook" for beginners that describes the use of computer programs was written by Hall (2001).

The interested user of software can study tutorials available in the **internet**. The addresses of these can be found with search engines using relevant keywords. Some examples:

B.R. Spear et al. (1995 ff.), Introduction to cladistics:
http://www.ucmp.berkeley.edu/clad/clad1.html

A. Dress (1995), Mathematical foundations of molecular systematics:
http://www.biotech.ist.unige.it/bcd/Curric/MathAn/mathan.html

Reconstruction of DNA secondary structure:
http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi
http://www.rna.icmb.utexas.edu/
http://wwwbio.leidenuniv.nl/~Batenburg/STRAbout.html

Alignment methods:
http://www.ridom-rdna.de/qalign/
http://bioweb.pasteur.fr/seqanal/interfaces/clustalw-simple.html
http://www.mbio.ncsu.edu/BioEdit/bioedit.html
ftp://ftp.amnh.org/pub/molecular/poy

Alignment using secondary structure information:
http://plaza.snu.ac.kr/~jchun/phydit/

Simulation of sequence evolution and tools for sequence analyses:
http://evolve.zoo.ox.ac.uk/

Drawing and printing trees from tree files with TREEVIEW:
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Furthermore there are lists of servers which offer computer programs
http://www.ability.org.uk/biomath.html

Overview on available computer programs for phylogenetic analyses (with addresses where they can be obtained):
http://evolution.genetics.washington.edu/phylip/software.html

Minimum evolution method:
http://www.lirmm.fr/~w3ifa/MAAS/FastME/FastME.html

Search for information in gene databases:
http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html

Simulations of genetic processes within populations:
http://darwin.eeb.uconn.edu/simulations/simulations.html

Population genetic analyses:
http://lgb.unige.ch/arlequin/

PAUP can be ordered from Sinauer Associates, Sunderland:
http://www.sinauer.com

As internet sites age very rapidly, it is recommended to search for the names of methods with the help of search engines (e.g., "multiple alignment", "maximum likelihood" or "split decomposition") or using names of computer programs.

16. References

Adachi, J., Cao, Y. & Hasegawa, M. (1993): Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid level: rapid evolution in warm-blooded vertebrates. – J. Mol. Evol. 36: 270-281.

Adachi, J. & Hasegawa, M. (1992): MOLPHY – Programs for molecular phylogenetics I – PROTML: Maximum likelihood inferrence of protein phylogeny. – Coputer Science Monographs 27, Institute of Statistical Mathematics, Tokyo.

Adell, J. C. & Dopazo, J. (1994): Monte Carlo simulation in phylogenies: An application to test the constancy of evolutionary rates. – J. Mol. Evol. 38: 305-309.

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M., Garey, J. R., Raff, R. A. & Lake, J. A. (1997): Evidence for a clade of nematodes, arthropods and other moulting animals. – Nature 387: 489-492.

Albert V. A., Mishler B. D. & Chase M. W. (1992): Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. – In: Soltis P. S., Soltis D. E. & Doyle J. J. (eds.), Molceular systematics of plants, Chapman & Hall, New York: 369-401.

Altschul, S. F. (1991): Amino acid substitution matrices from an information theoretic perspective. – J. Mol. Biol. 219: 555-565.

Alzogaray, R. A. (1998): Molecular basis of insecticide resistance. – Acta Bioquim. Clinica Latinoam. 32: 387.

Anderson D. T. (1967): Larval development and segment formation in the branchiopod crustaceans *Limnadia stanleyana* King (Conchostraca) and *Artemia salina* (L.) (Anostraca). – Aust. J. Zool. 15: 47-91.

Archie, J. W. (1989): Homoplasy excess ratio: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. – Syst. Zool. 38: 253-269.

– (1996): Measures of homoplasy. – In: Sanderson, M. J. & Hufford, L. (eds.), Homoplasy – the recurrence of similarity in evolution. Academic Press, San Diego, 153-188.

Aris-Brosou, S., Yang, Z. (2002): Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. – Syst. Biol. 51: 703-714.

Àrnason, U. & Gullberg, A. (1993): Comparison between the complete mtDNA sequence of the blue and the fin whale, two species that can hybridize in nature. – J. Mol. Evol. 37: 312-322.

Àrnason, U., Gullberg, A., Johnsson, E. & Ledje, C. (1993): The nucleotide sequence of the mitochondrial DNA molecule of the gray seal, *Halichoerus grypus*, and a comparison with mitochondrial sequences of other true seals. – J. Mol. Evol. 37: 323-330.

Arnold, M. L. (1997): Natural hybridization and evolution. – Oxford Univ. Press, New York.

Asakawa S., Kumazawa Y., Araki T., Himeno H., Miura K. & Watanabe K. (1991): Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. – J. Mol. Evol. 32: 511-520.

Ashe, J. S. & Marx, H. (1990): Phylogeny of viperine snakes (Viperinae): Part II. Cladistic analysis and major lineages. – Fieldiana Zool. N.S. 52: 1-23.

Attenborough, D. (1998): The life of birds. – BBC Books, London.

Averof, M. & Cohen S. M. (1997): Evolutionary origin of insect wings from ancestral gills. – Nature 385: 627-630

Ax, P. (1984): Das Phylogenetische System. – G. Fischer Verlag, Stuttgart.

– (1987): The phylogenetic system. The systematization of organisms on the basis of their phylogenies. – J. Wiley, Chichester.

– (1988): Systematik in der Biologie. – UTB – G. Fischer, Stuttgart.

– (1995): Das System der Metazoa I. G. – Fischer Verlag, Stuttgart.

– (1999): Das System der Metazoa II. – G. Fischer Verlag, Stuttgart.

Bachmann, K. (1998): Species as units of diversity: an outdated concept. – Theory Biosci. 117: 213-231.

Ballard J. W. O., Olsen G. J., Faith D. P., Odgers W. A., Rowell D. M. & Atkinson P. W. (1992): Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. – Science 258: 1345-1348.

Ballard J. W. O. & Kreitman M. (1995): Is mitochondrial DNA a strictly neutral marker? TREE 10:485-488

Bandelt, H. J. (1994): Phylogenetic networks. – Verh. Naturwiss. Ver. Hamburg 34: 51-71

Bandelt, H. J. & Dress, A. (1992): A new and useful approach to phylogenetic analysis of distance data. – Mol. Phylog. Evol. 1: 242-252.

– (1993): A relational approach to split decomposition. – In: Opitz, O., Lausen, B. & Klar, R. (eds.), Information and Classification. Springer Verlag, Berlin: 123-131.

Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. (1995): Mitochondrial portrait of human populations using median networks. – Genetics 141: 743-753.

Bandelt, H. J., Forster, P. & Röhl, A. (1999): Medina-joining networks for inferring intraspecific phylogenies. – Mol. Biol. Evol. 16: 37-48.

Bandelt, H. J., Macaulay, V. & Richards, M. (2000): Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. – Mol. Phylog. Evol. 16: 8-28.

Barnard J. L. & Ingram C. (1990): Lysianassoid Amphipoda (Crustacea) from deep-sea thermal vents. – Smiths. Contrib. Zool. 499: 1-80.

Baum, B. R. (1992): Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. – Taxon 41: 3-10.

Benecke, M. (1998): Random amplified polymorphic DNA (RAPD) typing of necrophageous insects (diptera, coleoptera) in criminal forensic studies: validation and use in practice. – Forensic Sci. Intern. 98: 157-168

Benton, M. J. (2000): Stems, nodes, crownclades, and rank-free lists: is Linnaeus dead? – Biol. Rev. 75: 633-648.

Berbee, M. L. & Taylor, J. W. (1993): Dating the evolutionary radiations of the true fungi. – Can. J. Bot. 71: 1114-1127.

Berg, D. E., Akopyants, N. S. & Kersulyte, D. (1994): Fingerprinting microbial genomes using the RAPD or AP-PCR method. – Methods Mol. Cell. Biol. 5: 13-24.

Bharathan G., Janssen B.-J., Kellogg E. A. & Sinha N. (1997): Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? – Proc. Natl. Acad. Sci. USA 94: 13749-13753.

Bininda-Emonds, O. R. P. & Sanderson, M. J. (2001): Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. – Syst. Biol. 50: 565-579.

Blanchard, J. L. & Schmidt, G. W. (1995): Pervasive migration of organellar DNA to the nucleus in plants. – J. Mol. Evol. 41: 397-406.

Blanchette, M., Bourque, G. & Sankoff, D. (1997) : Breakpoint phylogenies. – In: Miyano, S. & Takagi, T. (eds.), Genome informatics. University Academy Press, Tokyo: 25-34.

Bleiweiss, R. (1997): Slow rate of molecular evolution in high-elevation hummingbirds. – Proc. Natl. Acad. Sci. USA 95: 612-616.

Boore J. L., Collins T. M., Stanton D., Daehler L. L. & Brown W. M. (1995): Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. – Nature 376:163-165.

Boursot, P., Din, W., Anand, R., Darviche, D., Dod, R., Von Deimling, F., Talwar, G. P. & Bonhomme, F. (1996): Origin and radiation of the house mouse: mitochondrial DNA phylogeny. – J. Evol. Biol. 9: 391-415.

Bowles, J., Hope, M., Tiu, W. U., Liu, X. & McManus, D. P. (1993): Nuclear and mitochondrial genetic markers highly conserved between Chinese and Philippine *Schistosoma japonicum*. – Acta Tropica 55: 217-229.

Bremer, K. (1988): The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. – Evolution 42: 795-803.

Briggs D. E. G. (1992): Phylogenetic significance of the Burgess Shale crustacean *Canadaspis*. – Acta Zool 73: 293-300.

Briggs D. E. G., Fortey R. A. & Wills M. A. (1992): Morphological disparity in the Cambrian. – Science 256: 1670-1673.

Briggs, D. E. G., Erwin, D. H. & Collier, F. J. (1994): The fossils of the Burgess Shale. – Smiths. Inst. Press, Wahsington & London.

Brock, T. D., Madigan, M. T., Martinko, J. M. & Parker, J. (1994): Biology of microrogranisms, 7. Auflage. – Prentice Hall Inc.

Bromham L., Rambaut A. & Harvey P. H. (1996): Determinants of rate variation in mammalian DNA sequence evolution. – J. Mol. Evol. 43: 610-621.

Bromham, L. & Penny, D. (2003): The modern molecular clock. – Nature Rev. Gen. 4: 216-224.

Brooks, D. R., O'Grady, R. T. & Wiley, E. O. (1986): A measure of the information content of phylogenetic trees, and its as an optimality criterion. – Syst. Zool. 35: 571-581.

Brower, A. V. Z. & DeSalle, R. (1994): Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. – Ann. Entomol. Soc. Am. 87: 702-716.

Brusca, R. C. & Brusca, G. J. (2003): Invertebrates (second ed.). – Sinauer Assoc. Inc., Sunderland.

Bryant, H. N. (1989): An evaluation of cladistic and character analyses as hypothetico-deductive procedures, and the consequences for character weighting. – Syst. Zool. 38: 214-227.

Bryant, D. & Moulton, V. (2002): NeighbourNet: an agglomerative method for the construction of planar phylogenetic trees. – In: Guigo, R. & Gusfield, D. (eds.), Workshop in algorithms for bioinformatics. Springer Verlag, Heidelberg: 375-391.

Buckley, T. R., Simon, C. & Chambers, G. K. (2001): Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. – Syst. Biol. 50: 67-86.

Buckley, T. R., Arensburger, P., Simon, C. & Chambers, G. (2002): Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. – Syst. Biol. 51: 4-18.

Bulmer, M. (1988): Are codon usage patterns in unicellular organisms determined by selection-mutation balance? – J. Evol. Biol. 1: 15-26.

Buneman, P. (1971): The recovery of trees from measures of dissimilarity. – In: Hodson, F. R.; Kendall, D. G. & Tautu, P. (eds.), Mathematics in the archeological and historical sciences. Edinburgh Univ. Press, Edinburgh: 387-395.

Bunge, M. (1967): Scientific Research I. The search for system. – Springer-Verlag, Berlin.

Buntjer, J. B., Hoff, I. A. & Lenstra, J. A. (1997): Artiodactyl interspersed DNA repeats in Cetacean genomes. – J. Mol. Evol. 45: 66-69.

Burow, M. D., Simpson, C. E., Paterson, A. H. & J. L. Starr (1996): Identification of peanut (*Arachishypogaea* L.) RAPD markers diagnostic of root-knot nematode (*Meloidogyne arenaria* (Neal) Chitwood) resistance. – Mol. Breeding 2: 369-379.

Cadle, J. E. (1988): Phylogenetic relationships among advanced snakes. A molecular perspective. – Univ. Calif. Publ. 119: 1-77

Camin, J. H. & Sokal, R. R. (1965): A method for deducing branching sequences in phylogeny. – Evolution 19: 311-326.

Cantino, P. D., Bryant, H. N., de Queiroz, K., Donoghue, M. J., Eriksson, T., Jillis, D. M. & Lee, M. S. Y. (1999): Species names in phylogenetic nomenclature. – Syst. Biol. 48: 790-807.

Carroll, R. L. (1993): Paläontologie und Evolution der Wirbeltiere. – G. Thieme Verlag, Stuttgart.

Carroll, S. B. (1994): Developmental regulatory mechanisms in the evolution of insect diversity. – Development Suppl. 1994: 217-223.

Carroll, S. B., Grenier, J. K. & Weatherbee, S. B. (2001): From DNA to diversity. – Blackwell Science, Malden (USA) and Oxford (England).

Charleston, M. A., Hendy, M. D. & Penny, D. (1994): The effects of sequence length, tree topology and number of taxa on the performance of phylogenetic methods. – J. Comp. Biol. 1: 133-151.

Chippindale, P. T. & Wiens, J. J. (1994): Weighting, partitioning, and combining characters in phylogenetic analysis. – Syst. Biol. 43: 278-287.

Cifelli R. L. (1993): Theria of Metatherian-Eutherian grade and the origin of marsupials. – In: Szalay F. S. & Novaceck M. J., (eds.), Mammal phylogeny. Springer Verlag, Berlin, Heidelberg, New York: 205-215.

Clark, J. B., Maddison, W. P. & Kidwell, M. G. (1994): Phylogenetic analysis supports horizontal transfer of P transposable elements. – Mol. Biol. Evol. 11: 40-50.

Claverie, J. M. (1993): Detecting frame shifts by amino acid sequence comparison. – J. Mol. Biol. 234: 1140-1157.

Clement, P., Harris, A. & Davis, J. (1993): Finches and Sparrows. An identification guide. – Christopher Helm Ltd., London.

Cohen, S. & Jürgens, G. (1991): *Drosophila* headlines. – Trends Genetics 7: 267-271.

Cover, T. M. & Thomas, J. A. (1991): Elements of Information Theory, vol 1. – J. Wiley & Sons, New York.

Cracraft, J. (1987): Species concepts and the ontology of evolution. – Biol. Philos. 2: 329-346.

– (1989): Speciation and its ontology: The empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. – In: Otte, D. & Endler, J. A. (eds.), Speciation and its consequences. Sinauer, Sunderland: 28-59.

Craw, R. (1992): Margins of cladistics: identity, difference and place in the emergence of phylogenetic systematics, 1864-1975. – In: Griffiths, P. E. (ed.), Trees of life: essays in the philosophy of biology. Dordrecht, pp. 65-107.

Crochet, P. A., Dubois, A., Ohler, A. & Tunner, H. (1995): *Rana (Pelophylax) ridibunda* Pallas, 1771, *Rana (Pelophylax) perezi* Seoane, 1885 and their associated kleptons (Amphibia, Anura): morphological diagnoses and description of a new taxon. – Bull. Mus. nat. Hist. nat., Paris, (4) **17** A (1-2): 11-30.

Crozier, R. H. & Crozier, Y. C. (1993): The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. – Genetics 133: 97-117.

Cunningham, C. W. (1997): Can three incongruence tests predict when data should be combined? – Mol. Biol. Evol. 14: 733-740.

Danielopol, D. L. (1977): On the origin and diversity of European freshwater interstitial Ostracods. – 6th Intern. Ostracod Symp.: 295-305.

Darwin, C. (1859): On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. – J. Murray, London

Davison, D. (1985): Sequence similarity ("homology") searching for molecular biologists. – Bull. Math. Biol. 47: 437-474.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978): A model of evolutionary change in proteins. – In: Dayhoff, M. O. (ed.), Atlas of protein sequence and structure, Vol. 5. National Biomedical Research Foundation, Silver Spring: 89-99.

De Jong, R. (1980): Some tools for evolutionary and phylogenetic studies. – Z. zool Syst. Evol.-forsch. 18: 1-23.

Dendy, A. (1912): Outlines of evolutionary biology. – Constable & Co. Ltd., London.

De Pinna, M. C. C. (1991): Concepts and tests of homology in the cladistic paradigm. – Cladistics 7: 367-394.

De Queiroz, K. (1985): The ontogenetic method for determining character polarity and its relevance to phylogenetic systematics. – Syst. Zool. 34: 280-299.

De Rijk, P. & De Wachter, R. (1993): DCSE, an interactive tool for sequence alignment and secondary structure research. – CABIOS 9: 735-740.

De Robertis, E. M. (1997): The ancestry of segmentation. – Nature 387:25-26.

Disotell, T. R. R., Honeycutt, R. L. & Ruvolo, M. (1992): Mitochondrial DNA phylogeny of the old-world monkey tribe Papionini. – Mol. Biol. Evol. 9: 1-13.

Desper, R. & Gascuel, O., (2002): Fast and accurate phylogeny reconstruction with algorithms based on the minimum-evolution principle. – In: Guigo, R., Gusfield, D. (eds.), Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI'2002, Roma). Lecture Notes in Computer Science 2452: 357-374.

Dobzhansky, T. & Spassky, B. (1959): *Drosophila paulistorum*, a cluster of species in statu nascendi. – Proc Natl Acad Sci 45:419-428.

Dollo, L. (1893): Les lois de l'évolution. – Bull. Soc. Belge Geol. Paléont. Hydrol. 7: 164-166.

Dowling, T. E., Moritz, C., Palmer, J. D. (1990): Nucleic acids II: restriction site analysis. – In: Hillis, D. M. & Moritz, C. (eds.), Molecular Systematics. – Sinauer Ass., Sunderland: 250-317.

Downie, S. R., Olmstead, R. G., Zurawski, G., Soltis, D. E., Soltis, P. S., Watson, J. C. & Palmer, J. D. (1991): Six independent losses of the chloroplast DNA *rpl2* intron in dicotyledons: molecular and phylogenetic implications. – Evolution 45: 1245-1259.

Dress, A. (1995): The mathematical basis of molecular phylogenetics. http://www.techfak.uni-bielefeld.de/bcd/Curric/MathAn/node2.html.

Dreyer, H. & Wägele, J.-W. (2001): Parasites of crustaceans (Isopoda: Bopyridae) evolved from fish parasites: Molecular and morphological evidence. – Zoology 103: 157-178.

Ebeling, W. (1990): Physik der Evolutionsprozesse. – Akademie-Verlag, Berlin.

Eck, R. V. & Dayhoff, M. O. (1966): Atlas of protein sequence and structure. – Natl. Biomed. Res. Found., Silver Springs, Maryland.

Ehlers, U. (1985): Das phylogenetische System der Plathelminthes. – G. Fischer Verlag, Stuttgart.

Edwards, A. W. F. & Cavalli-Sforza, L. L. (1963): The reconstruction of evolution. – Ann. Hum. Genet. 27: 104-105.

Edwards, A. W. F. (1996): The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. – Syst. Biol. 45: 79-91.

Elder J. F. & Turner B. J. (1995): Concerted evolution of repetitive DNA sequences in eukaryotes. – Quart. Rev. Biol .70:297-320.

Emerson, S. B. (1998): Morphological correlations in evolution: consequences for phylogenetic analysis. – Quart. Rev. Biol. 73: 141-162.

Estabrook, G. F., Johnson, C. S. Jr. & McMorris, F. R (1976a): A mathematical foundation for the analysis of character compatibility. – Mathematical Biosciences 23: 181-187.

– (1976b): An algebraic analysis of cladistic characters. – Discrete Mathematics 16: 141-147.

Estabrook, G. F., Strauch, G. F. & Fiala, K. (1977): An application of compatibility analysis to Blacklith's data on orthopteroid insects. – Syst. Zool. 26: 269-276.

Faith, D. P. (1991): Cladistic permutation tests for monohyly and nonmonophyly. – Syst. Zool. 40: 366-375.

Faith, D. P., Cranston, P. S. (1991): Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. – Cladistics 7: 1-28.

Farris, J. S. (1969): A successive approximations approach to character weighting. – Syst. Zool. 18: 374-385.

– (1970): Methods for computing Wagner trees. – Syst. Zool. 19: 83-92.

– (1972): Estimating phylogenetic trees from distance matrices. – Am. Nat. 106: 645-668.

– (1982): The logical basis of phylogenetic analysis. – In: Platnick, N. & Funk, V. (eds.), Advances in Cladistics 2. Columbia Univ. Press, New York: 7-36.

– (1989a): The retention index and homoplasy excess. – Syst. Zool. 38: 406-407.

– (1989b): The retention index and the rescaled consistency index. – Cladistics 5: 417-419.

– (1991): Excess homoplasy ratios. – Cladistics 7: 81-91.

Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D. & Kluge, A. G. (1996): Parsimony jackknifing outperforms neighbour-joining. – Cladistics 12: 99-124

Farris, S. J., Källersjö, M., Kluge, A. G. & Bult, C. (1995): Constructing a significance test for incongruence. – Syst. Biol. 44: 570-572.

Felsenstein, J. (1978a): The number of evolutionary trees. – Syst. Zool. 27: 27-33.

– (1978b): Cases in which parsimony and compatibility methods will be positively misleading. – Syst. Zool. 27: 401-410.

– (1981): Evolutionary trees from DNA sequences: a maximum likelihood approach. – J. Mol. Evol. 17: 368-376.

– (1983): Statistical inference of phylogenies. – J. Roy. Statist. Soc. A 146: 246-272.

– (1984): Distance methods for inferring phylogenies: A justification. – Evolution 38: 16-24.

– (1985): Confidence limits on phylogenies: an approach using the bootstrap. – Evolution 39: 783-791.

– (1991): Counting phylogenetic invariants in some simple cases. – J. Theoret. Biol. 152: 357-376.

– (1993): PHYLIP (Phylogeny Inference Package) version 3.5c. – Distributed by the author. Department of Genetics, University of Washington, Seattle.

Ferrière, R. & Fox, G. A. (1995): Chaos and evolution. – TREE 10: 480-485

Fitch, W. M. (1967): Construction of phylogenetic trees. – Science 155: 279-284.

– (1971): Toward defining the course of evolution: minimum change for a specified tree topology. – Syst. Zool. 20: 406-416.

– (1984): Cladistic and other methods: problems, pitfalls, and potentials. – In: Duncan, T. & Stuessy, T. F. (eds.), Cladistics: perspectives on the reconstruction of evolutionary history. Columbia Univ. Press, New York: 221-252.

Fitch, W. M. & Margoliash, E. (1967): Construction of phylogenetic trees. – Science 155: 279-284.

Fitch, W. M. & Ye, J. (1991): Weighted parsimony: does it work? – In: Miyamoto, M. M. & Cracraft, J. (eds.), Phylogenetic analysis of DNA sequences. Oxford University Press, New York, 147-154.

Friedrich, M. & Tautz, D. (1995) Ribosomal DNA phylogeny of the major extant classes and the evolution of myriapods. – Nature 376: 165-167.

– (1997): An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. – Mol. Biol. Evol. 14:644-653.

Fu, Y. X. & Li, W. H. (1993): Statistical tests of neutrality of mutations. – Genetics 133: 693-709.

Funch, P. & Kristensen, R. M. (1995): Cyclicphora is a new phylum with affinities to Entoprocta and Ectoprocta. – Nature 378: 711-714.

Garcia-Vallve, S., Romeu, A. & Palau, J. (2000): Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. – Mol. Biol. Evol. 17: 352-361.

Gascuel, O., Bryant, D. & Denis, F. (2001): Strengths and limitations of the minimum-evolution principle. – Syst. Biol. 50: 621-627.

Gatesy, J. E., O'Grady, P. M. & Baker, R. H. (1999): Corroboration among datasets in simultaneous analysis: hidden support for phylogenetic relationships among higher-level artiodactyl taxa. – Cladistics 15: 271-313.

Gaunt, S. J. (1997): Chick limbs, fly wings and homology at the fringe. – Nature 386:324-325.

Geller, J. B. (1994): Sex-specific mitochondrial DNA haplotypes and heteroplasmy in *Mytilus trossulus* and *Mytilus galloprovincialis* populations. – Mol. Mar. Biol. Biotechnol. 3(6): 334-337.

Ghiselin, M. T. (1966): On psychologism in the logic of taxonomic controversies. – Syst. Zool. 15: 207-215.

– (1974): A radical solution to the species problem. – Syst. Zool. 23: 536-544.

Gibbons, A. (1998): Calibrating the molecular clock. – Science 279: 28-29.

Givnish, T. J. & Sytsma, K. J. (1997): Homoplasy in molecular vs. morphological data: the likelihood of correct phylogenetic inference. – In: Givnish, T. J. & Sytsma, K. J. (eds.), Molecular evolution and adaptive radiation, 1st edn. Cambridge Univ. Press, Cambridge U.K.: 55-101.

Goethe, J. W. (1824): Zahme Xenien, Buch III. [In: E. Beutler, 1977 (ed.), Johann-Wolfgang Goethe – Sämtliche Werke. Band I. Sämtliche Gedichte. – Nachdruck der Artemis-Gedenkausgabe von 1949. Artemis Verlags-AG, Zürich: 629].

Gogarten, J. P. (1995): The early evolution of cellular life. – TREE 10: 147-151.

Gojobori, T., Li, W. H. & Graur, D. (1982): Patterns of nucleotide substitution in pseudogenes and functional genes. – J. Mol. Evol. 18: 360-369.

Goldmann, N. (1993): Simple diagnostic statistical test of models for DNA substitution. – J. Mol. Evol. 37: 650-661.

– (1993): Statistical tests of models of DNA substitution. – J. Mol. Evol. 36: 182-198.

Goldman, N., Anderson, J. P. & Rodrigo, A. G. (2000): Likelihood-based tests of topologies in phylogenetics. – Syst. Biol. 49: 652-670.

Goloboff, P. A. (1991): Homoplasy and the choice among cladograms. – Cladistics 7: 215-232.

– (1993): Estimating character weights during tree search. – Cladistics 9: 83-91.

– (2001): Techniques for analyzing large datasets. – In: DeSalle, R., Giribet, G. & Wheeler, W.: Techniques in molecular systematics and evolution. Birkhäuser Verlag, Basel: 70-80.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992): Exhaustive matching of the entire protein sequence database. – Science 256: 1443-1445.

Goodman, M., Bailey, W. J., Hayasaka, K., Stanhope, M. J., Slightom, J. & Czelusniak, J. (1994): Molecular evidence on primate phylogeny from DNA sequences. – Am. J. Phys. Anthropol. 94: 3-24.

Gould, S. J. (1989): Wonderful life, the Burgess shale and the nature of history. – W.W. Norton, New York.

Graham, R. L. & Foulds, L. R. (1982): Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. – Mathem. Biosci. 60: 133-142.

Grant, P. R. (1993): Hybridization of Darwin's finches on Isla Daphne Major, Galápagos. – Phil. Trans. R. Soc. Lond. B 340: 127-139.

Grant, V. (1994): Evolution of the species concept. – Biol. Zentralbl .113: 401-415.

Grantham, R., Gauthier, C., Gouy, R., Mercier, R. & Pave, A. (1980): Codon catalog usage and the genome hypothesis. – Nucleic Acid Res. 8: r49-r62.

Green, S. & Chambon, P. (1986): A superfamily of potentially oncogenic hormone receptors. – Nature 324: 615-617.

Greenwood, J. J. D. (1993): Theory fits the bill in the Galápagos Islands. – Nature 362: 699

Grishin, N. V. (1995): Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. – J. Mol. Evol. 41: 675-679.

Grosberg, R. K., Levitan, D. R. & Cameron, B. B. (1996): Characterization of genetic structure and genealogies using RAPD-PCR markers: a random primer for the novice and nervous. – In: Ferraris, J. D. & Palumbi, S. R., Molecular Zoology: Advances, strategies, and protocols: 67-100.

Gu, W. & Li, W. H. (1992): Higher rates of amino acid substitutions in rodents than in humans. – Mol. Phylog. Evol. 1: 211-214.

Haeckel, E. (1866): Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformierte Descendenz-Theorie. – Georg Reimer Verlag, Berlin.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastes, J. W. & Nadler, S. A. (1994): Disparate rates of molecular evolution in cospeciating hosts and parasites. – Science 265: 1087-1090.

Halanych, K. M. (1995): The phylogenetic position of the pterobranch hemichordates based on 18SrDNA sequence data. – Mol. Phylog. Evol. 4: 72-76.

Hall, B. G. (2001): Phylogenetic trees made easy. – Sinauer Assoc., Sunderland.

– (1992): Evolutionary developmental biology. – Chapman & Hall, London, 1-275.

Harrington, R. W. & Kallman, K. D. (1968): The homozygosity of clones of the self-fertilizing hermaphroditic fish *Rivulus marmoratus* (Cyprinodontidae, Atheriniformes). – Amer. Nat. 102: 337-343.

Harshman, J. (1994): The effect of irrelevant characters on bootstrap values. – Syst. Biol. 43: 419-424.

Hasegawa, M., Kishino, H. & Yano, T. (1985): Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. – J. Mol. Evol. 22: 160-174.

Hasegawa M., Kishino H. & Saitou N. (1991): On the maximum likelihood method in molecular phylogenetics. – J. Mol. Evol. 32: 443-445.

Hassenstein, B. (1966): Was ist Information? – Naturwiss. Medizin 3: 38-52.

Hastings, W. (1970): Monte Carlo sampling methods using Markov chains and their applications. – Biometrika 57: 97-109.

Hendy, M. D. & Penny, D. (1989): A framework for the quantitative study of evolutionary trees. – Syst. Zool. 38: 297-309.

– (1993): Spectral analysis of phylogenetic data. – J. Classif. 10: 5-24.

Hendy, M. D., Penny, D. & Steel, M. A. (1994): A discrete Fourier analysis for evolutionary trees. – Proc. Natl. Acad. Sci. USA 91: 3339-3343.

Henikoff, S. & Henikoff, J. G. (1992): Amino acid substitution matrices from protein blocks. – Proc. Nat. Acad. Sci. 89*:* 10915-10919.

Hennig, W. (1949): Zur Klärung einiger Begriffe der phylogenetischen Systematik. – Forsch. Fortschr. 25: 136-138.

– (1950): Grundzüge einer Theorie der phylogenetischen Systematik. – Deutscher Zentralverlag Berlin.

– (1953): Kritische Bemerkungen zum phylogenetischen System der Insekten. – Beitr. Entomol. 3: 1-61.

– (1966): Phylogenetic Systematics. – Univ. Illinois Press, Urbana.

– (1982): Phylogenetische Systematik. – Paul Parey, Berlin und Hamburg, 1-246.

– (1986): Taschenbuch der speziellen Zoologie, Wirbellose 2: Gliedertiere. – Harri Deutsch, Thun, Frankfurt.

Hessler, R. R. & Martin, J. W. (1989): *Austinograea williamsi*, new genus, new species, a hydrothermal vent crab (Decapoda: Bythograeidae) from the Mariana back-arc basin, Western Pacific. – J. Crust. Biol. 9: 645-661.

Higgins, D. G. & Sharp, P. M. (1988): CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. – Gene 73: 237-244.

Hillis, D. M. (1984): Misuse and modification of Nei's genetic distance. – Syst. Zool. 33: 238-240.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. & Molineux, I. J. (1992): Experimental phylogenetics: generation of a known phylogeny. – Science 255: 589-592.

Hillis, D. M. & Huelsenbeck, J. P. (1992): Signal, noise, and reliability in molecular phylogenetic analyses. – J. Hered. 83: 189-195.

Hillis, D. M., Mable, B. K. & Moritz, C. (1996): Applications of molecular systematics: the state of the field and a look into the future. – In: Hillis, D. M., Moritz, C. & Mable, B. K. (eds.), Molecular Systematics. Sinauer Ass., Sunderland: 515-543.

Hoyle, D. C. & Higgs, P. G. (2003): Factors affecting errors in the estimation of evolutionary distances between sequences. – Mol. Biol. Evol. 20: 1-9.

Hudson, R. R. (1993): The how and why of generating gene genealogies. – In: Takahata, N. & Clark, A.G. (eds.), Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology. Sinauer Ass., Sunderland: 23-36.

Huelsenbeck, J. P. & Bull, J. J. (1996): A likelihood ratio test to detect conflicting phylogenetic signal. – Syst. Biol. 45: 92-98.

Huelsenbeck, J. P., Bull, J. J. & Cunningham, C. W. (1996): Combining data in phylogenetic analysis. – TREE 11: 152-158.

Huelsenbeck, J. P., Hillis, D. M. & Jones, R. (1996b): Parametric bootstrapping in molecular phylogenetics: applications and performance. – In: Ferraris, J. D. & Palumbi, S. R., Molecular Zoology: Advances, strategies, and protocols: 19-45.

Huelsenbeck, J. P. & Ronquist, F. (2001): MrBAYES: Bayesian inference of phylogenetic trees. – Bioinformatics 17: 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001): Bayesian inference of phylogeny and its impact on evolutionary biology. – Science 294: 2310-2314.

Huey, R. B. & Bennett, A. F. (1987): Phylogenetic studies of coadaptation: preferred temperatures versus optimal performance temperatures of lizards. – Evolution 41: 1098-1115.

Hughes, A. L. (1994): Evolution of the interleukin-1 gene family in mammals. – J. Mol. Evol. 39:6-12.

Hughes, A. L. & Yeager, M. (1997): Comparative evolutionary rates of introns and exons in murine rodents. – J. Mol. Evol. 45: 125-130.

Hume, D. (1777): Enquiries concerning human understanding and concerning the principles of morals. Reprinted 1996 from the posthumous edition of 1777. – Clarendon Press, Oxford.

Huys, R. & Boxshall, G. A. (1991): Copepod evolution. – The Ray Society, London.

Hwang, U. W., Kim, W., Tautz, D. & Friedrich, M. (1998): Molecular phylogenetics at the Felsenstein Zone: approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. – Mol. Phylog. Evol. 9: 470-480.

Ichinose, H., Ohye, T., Takahashi, E., Seki, N., Hori, T., Segawa, M., Nomura, Y., Endo, K., Tanaka, H., Tsuji, S., Fujita, K. & Nagatsu, T. (1994): Hereditary progressive dystonia with marked diurnal fluctuation caused by mutations in the GTP cyclohydrolase I gene. – Nature Genetics 8: 236-242.

International code of nomenclature of bacteria (1992). – American Society for Microbiology, Washington D.C.

International code of botanical nomenclature (1994). – Koeltz Scientific Books, Königstein, Germany.

International code of zoological nomenclature (1999). – International Trust for Zoological Nomenclature, London.

Irwin, D. M., Kocher, T. D. & Wilson, A. C. (1991): Evolution of the cytochrome b gene of mammals. – J. Mol. Evol. 32: 128-144.

Janich, P. (1997): Kleine Philosophie der Naturwissenschaften. – Beck'sche Verlagsbuchhandlung, München

Janke, A., Gemmell, N. J., Feldmaier-Fuchs, G., Haeseler, A. von & Pääbo, S. (1996): The mitochondrial genome of a monotreme – the platypus *(Ornithorhynchus anatinus)*. – J. Mol. Evol. 42: 153-159.

Jefferies, R. P. S (1979): The origin of chordates – a methodological essay. – In: House, M. R. (ed.), The origin of the major vertebrate groups. Academic Press, London: 443-477.

Jenner, R. A. (2002): Boolean logic and character state identity: pitfalls of character coding in metazoan cladistics. – Contr. Zool. 71: 67-91.

Jin, L. & Nei, M. (1990): Limitations of the evolutionary parsimony method of phylogenetic analysis. – Mol. Biol. Evol. 7: 82-102.

Joger, U. (1996): Molekularbiologische Methoden in der phylogenetischen Rekonstruktion. – Zool. Beitr. N.F. 37: 77-131.

Johns, G. C. & Avise, J. C. (1998): A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. – Mol. Biol. Evol. 15: 1481-1490.

Johnson, C. (1986): Parthenogenetic reproduction in the philosciid isopod, *Ocelloscia floridana* (Van Name, 1940). – Crustaceana 51: 123-132.

Jukes, T. H. & Cantor, C. R. (1969): Evolution of protein molecules. – In: Munro, H. N. (ed.), mammalian protein Metabolism. Academic Press, New York: 21-132.

Jura, C. (1991): Phylogeny of Tracheata. – Przeglad Zoologiczny 34: 213-230.

Kaplan, N., Hudson, R. R. & Lizuka, M. (1991): The coalescent process in models with selection, recombination and geographic subdivision. – Genet. Res. 57: 83-91.

Kempken, F. (1995): Horizontal transfer of a mitochondrial plasmid. – Mol. Gen. Genet. 248: 89-94.

Killian, J. K., Buckley, T. R., Stewart, N., Munday, B. L. & Jirtle, R. L. (2001): Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. – Mammalian Genome 12: 513-517.

Kimura, M. (1962): On the probability of fixation of mutant genes in populations. – Genetics 47: 713-719.

– (1968): Evolutionary rate at the molecular level. – Nature 217: 624-626.

– (1980): A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. – J. Mol. Evol. 16: 111-120.

– (1981): Estimation of evolutionary distances between homologous nucleotide sequences. – Proc. Natl. Acad. Sci. USA 78: 454-458.

– (1983): The neutral theory of molecular evolution. – Cambridge University Press, Cambridge.

– (1987): Die Neutralitätstheorie der molekularen Evolution. – Verlag Paul Parey, Berlin.

Kimura, M. & Ohta, T. (1973): Mutation and evolution at the molecular level. – Genet. Suppl. 73: 19-35.

Kingman, J. F. C. (1982): On the genealogy of large populations. – J. Appl. Probab. 19A: 27-43.

Kishino, H. & Hasegawa, M. (1989): Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. – J. Mol. Evol. 29: 170-179.

Kluge, A. G. & Farris, J. S. (1969): Quantitative phyletics and the evolution of anurans. – Syst. Zool. 18: 1-32.

Knowlton, N., Weil, E., Weigt, L. A. & Guzmán, H. M. (1992): Sibling species in *Montastraea annularis*, coral bleaching, and the coral climate record. – Science 255: 330-333.

Koepcke, H. W. (1971-1973): Die Lebensformen. – Goecke & Evers, Krefeld.

Kollar, E. J. & Fisher, C. (1980): Tooth induction in chick epithelium: expression of quiescent genes for enamel synthesis. – Science 207: 993-995.

König, C. (1982): Zur systematischen Stellung der Neuweltgeier. – J. Ornithol. 123: 259-267.

Koenig, O. (1975): Biologie der Uniform. – In: von Ditfurth, H. (ed.): Evolution. Ein Querschnitt durch die Forschung. – Hoffmann & Campe, Hamburg: 175-211.

Kondo, R., Horai, S., Satta, Y. & Takahata, N. (1993): Evolution of hominid mitochondrial DNA with special reference to the silent substitution rate over the genome. – J. Mol. Evol. 36: 517-531.

Kraus, O. (1970): Internationale Regeln für die Zoologische Nomenklatur. – Waldemar Kramer Verlag, Frankfurt, 2. Auflage.

Kraus, O. & Kraus, M. (1994): Phylogenetic system of the Tracheata (Mandibulata): on "Myriapoda" – Insecta interrelationships, phylogenetic age and primary ecological niches. – Verh. Naturwiss. Ver. Hamburg 34: 5-31.

Kumar, S., Tamura, K. & Nei, M. (1993): MEGA: Molecular Evolutionary Genetics Analysis, Vers. 1.0. – The Pennsylvania State University, Univ. park, PA 16802.

Kumazawa, Y. & Nishida, M. (1995): Variations in mitochondrial tRNA gene organization of reptiles as phylogenetic markers. – Mol. Biol. Evol. 12: 759-772.

Kurtén, B. (1963): Return of a lost structure in the evolution of the felid dentition. – Soc. Sci. Fenn. Comment. Biol. 26: 1-12.

Kwiatowski, J., Krawczyk, M., Jaworski, M., Skarecky, D. & Ayala, F. J. (1997): Erratic evolution of glycerol-3-phosphate dehydrogenase in *Drosophila*, *Chymomyza*, and *Ceratitis*. – J. Mol. Evol. 44: 9-22.

Lake, J. A. (1987): A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. – Mol. Biol. Evol. 4: 167-191.

Lanave, C. G., Preparata, G., Saccone, C. & Serio, G. (1984): A new method for calculating evolutionary substitution rates. – J. Mol. Evol. 20: 86-93.

Lanyon, S. M. (1988): The stochastic mode of molecular evolution: what consequences for systematic investigations? – The Auk 105: 565-573.
Larget, B. & Simon, D. L. (1999): Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. – Mol. Biol. Evol. 16: 750-759.
Lauder, G. V. & Liem, K. F. (1983): The evolution and interrelationships of the actinopterygian fishes. – Bull. Mus. Comp. Zool. 150: 95-197.
Lauterbach, K. E. (1989): Das Pan-Monophylum – ein Hilfsmittel für die Praxis der phylogenetischen Systematik. – Zool. Anz. 223:139-156.
Lavin, M., Doyle, J. J. & Palmer, J. D. (1990): Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. – Evolution 44: 390-402.
Lento, G. M., Hickson, R. E., Chambers, G. K. & Penny, D. (1995): Use of spectral analysis to test hypotheses on the origin of pinnipeds. – Mol. Biol. Evol. 12: 28-52
Le Quesne, W. J. (1969): A method of selection of characters in numerical taxonomy. – Systematic Zoology 18: 201-205.
Lewis, P. O. (2001): Syst. Biol. 50: A likelihood approach to estimating phylogeny from discrete morphological character data. 913-925.
Lewontin, R. C. & Birch, L. C. (1966): Hybridization as a source of variation for adaptation to new environments. – Evolution, 20: 315-336.
Li, W. H. (1993): Unbiased estimation of the rates of synonymous and nonsynonymous substitution. – J. Mol. Evol. 36: 96-99.
– (1997): Molecular Evolution. – Sinauer Ass., Sunderland .
Li, W. H. & Graur, D. (1991): Fundamentals of molecular evolution. – Sinauer Ass., Sunderland.
Li, W. H., Wu, C. I. & Luo, C. C. (1984): Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. – J. Mol. Evol. 21: 58-71.
– (1985): A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. – Mol. Biol. Evol. 2: 150-174.
Linnaeus, C. (1758): Systema naturae per regna tria naturae, secundum classes, ordines, genera, species cum characteribus, differentiis, synonymis, locis. – Laurentii Salvii, Holmiae.
Lipman, D. J. & Pearson, W. R. (1985): Rapid and sensitive protein similarity searches. – Science 227: 1435-1441.
Liu H. P. & Mitton J. B. (1996): Tissue-specific maternal and paternal mitochondiral DNA in the freshwater mussel, *Anodonta grandis grandis*. – J. Moll. Stud. 62: 393-394.
Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994): Recovering evolutionary trees under a more general model of sequence evolution. – Mol. Biol. Evol. 11: 605-612.
Lopez, P., Forterre, P. & Philippe, H. (1999): A method for extracting ancient phylogenetic signal: the rooting of the universal tree of life based on elongation factors. – J. Mol. Evol. 49: 496-508.
Lorenz, K. (1941): Vergleichende Bewegungsstudien an Anatinen. J. Ornithol. Ergänzungsbd. 3: 194-293.
– (1941): Comparative studies on the behaviour of Anatinae. – Avicult. Mag. 59: 80-91.
– (1943): Psychologie und Stammesgeschichte. – In: Heberer, G., Die Evolution der Organismen: 105-127.
– (1973): Die Rückseite des Spiegels – Versuch einer Naturgeschichte menschlichen Erkennens. – Piper, München und Zürich.
Lutzoni, F., Wagner, P., Reeb, V. & Zoller, S. (2000): Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. – Syst. Biol. 49: 628-651.
Lyrholm, T., Leimar, O. & Gyllensten, U. (1990): Low diversity and biased substitution patterns in the mitochondrial DNA control region of sperm whales: Implications for estimates of time since common ancestry. – Mol. Biol. Evol. 13: 1318-1326.
Mabee, P. M. (1993): Phylogenetic interpretation of ontogenetic change: sorting out the actual and artefactual in an empirical case study of centrarchid fishes. – Zool. J. Linn. Soc. 107: 175-291.
Mac Arthur, R. H. & Wilson, E. O. (1967): The theory of island biogeography. – Princeton University Press, Princeton (New Jersey).
Macey J. R., Larson A., Ananjeva N. B. & Papenfuss T. J. (1997): Evolutionary shifts in three major structural features of the mitochondrial genome among iguanian lizards. – J. Mol. Evol. 44:660-674.
MacFadden, B. J. (1992): Fossil horses: systematics, paleobiology, and evolution of the family Equidae. – Cambridge Univ. Press, Cambridge.
Macgregor, H. C. & Varley, J. M. 1983: Working with chromosomes. – J. Wiley & Sons, Chichester.
Maddison W. P., Donoghue M. J. & Maddison D. R. (1984): Outgroup analysis and parsimony. – Systematic Zoology 33: 83-103.
Maddison, W. P. (1991): Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. – Syst. Zool. 40: 304-314.
Maddison, W. P., Maddison, D. R. (1992): MacClade Version 3. – Sinauer Assoc., Sunderland, Massachusetts.
Mahner, M. (1993): What is a species? A contribution to the never ending species debate in biology. – J. Gen. Philos. Sci. 24: 103-126.
Mahner, M. & Bunge, M. (1997): Foundations of biophilosophy. – Springer Verlag, Berlin.
Mai, J. C. & Coleman, A. W. (1997): The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. – J. Mol. Evol. 44: 258-271.
Marsh, O. C. (1880): *Odontornithes*: a monograph on the extinxt toothed birds of North America. – Rep. U.S. Geol. Explor. Fortieth Parallel, No. 7, Washington.

Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. & Kowallik, K. V. (1998): Gene transfer to the nucleus and the evolution of chloroplasts. – Nature 393: 162-165.

Mauersberger, G: (1974): Klasse Aves – Vögel, Rororo Tierwelt 4 (Vögel 1). – Rowohlt Verlag, Reinbeck bei Hamburg.

Maxson, L. R. (1992): Tempo and pattern in anuran speciation and phylogeny: An albumin perspective. – In: Adler, K. (ed.), Herpetology. Current research on the biology of amphibians and reptiles. Oxford (Ohio): 41-58.

Maxson, L. R. & Maxson R. D. (1990): Proteins II: Immunological techniques. – In: Hillis, D. M. & Moritz, C., Molecular Systematics. Sinauer Ass., Sunderland: 127-155.

Mayr, E. (1942): Systematics and the origin of species. – Columbia University Press, New York.

– (1963): Animal species and evolution. Harvard Univ. Press,Cambridge, Mass.

– (1969): Principles of Systematic Zoology. – MacGraw-Hill Book Co., New York

– (1981): Biological classification: Toward a synthesis of opposing methodologies. – Science 214: 510-516.

– (1982): The growth of biological thought. Diversity, evolution, and inheritance. – The Belcamp Press, Cambridge, Massachusetts.

– (1982): Speciation and macroevolution. – Evolution 36: 1119-1132.

Mayr, E. & Ashlock, P. D. (1991): Principles of systematic zoology; second edition. – McGraw Hill Inc., New York.

McDonald, J. H., Seed, R. & Koehn, R. K. (1991): Allozymes and morphometric characters of three species of *Mytilus* in the Northern and Southern Hemispheres. – Mar. Biol. 111: 323-333.

McLachlan, A. (1972): Repeating sequences and gene duplication in proteins. – J. Mol. Biol. 64: 417-437.

Meier, R. (1997): A test and review of the empirical performance of the ontogenetic criterion. – Syst. Biol. 46: 699-721.

Meinhard, H. (1997): Wie Schnecken sich in Schale werfen. Muster tropischer Meeresschnecken als dynamische Systeme. – Springer Verlag, Berlin.

Meinhardt, H. (1996): Models of biological pattern formation: common mechanism in plant and animal development. – Int. J. Dev. Biol. 40: 123-134.

Menzies, R. J. & George, R. Y. (1972): Isopod Crustacea of the Peru-Chile Trench. – Anton Bruun Rept. 9: 1-124.

Metropolois, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953): Equation of state calculations by fast computing machines. – J. Chem. Phys. 21: 1087-1092.

Meyrick, E. (1885): A Handbook of the British Lepidoptera. – MacMillan & Co., London.

Michener, C. D. (1977): Discordant evolution and the classification of Allodapine bees. – Syst Zool 26: 32-56

Mishler, B. & Brandon, R. (1987): Individuality, pluralism, and the phylogenetic species concept. – Biol. And Phil. 2: 397-414.

Mishler, B. D. (1994): Cladistic analysis of molecular and morphological data. – Amer. J. Phys. Anthropol. 94: 143-156.

Miyamoto, M. M. & Boyle, S. M. (1989): The potential importance of mitochondrial DNA sequence data to eutherian mammal phylogeny. – In: Fernholm, K., Bremer, K. & Jörnvall, H. (eds.), The Hierarchy of Life. Elsevier Sci. Publ.: 437-450.

Moore, R. C. (1969): Treatise on invertebrate paleontology, part R, Arthropoda. – Geological Society of America Inc., Lawrence, Kanada.

Moran, N. A., von Dohlen, C. D. & Baumann, P. (1995): Faster evolutionary rates in endosymbiotic bacteria than in cospeciating insect hosts. – J. Mol. Evol. 41: 727-731.

Morgenstern, B., Dress, A. & Werner, T. (1996): Multiple DNA and protein sequence alignment based on segment-to segment comparison. – Proc Natl Acad Sci USA 93: 12098-12103

Morrison, D. A. & Ellis, J. T. (1997): Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. – Mol. Biol. Evol. 14: 428-441.

Mow, W. H. (1994) Maximum likelihood sequence estimation from the lattice viewpoint. – IEEE Trans Inform Theory 40: 1591-1600.

Müller, F. (1864): Für Darwin. – Engelmann, Leipzig.

Müller, K. J. & Walossek, D. (1985): A remarkable arthropod fauna of the Upper Cambrian "Orsten" of Sweden. – Trans. Roy. Soc. Edinburgh 76: 161-172.

Müller, K. J. & Walossek, D. (1986): *Martinssonia elongata* gen. et sp.n., a crustacean-like euarthropod from the Upper Cambrian 'Orsten' of Sweden. – Zool. Scr. 15: 73-92.

Murphy, R. W., Sites, J. W., Buth, D. G. & Haufler, C. H. (1990): Proteins I: Isozyme electrophoresis. – In: Hillis, D. M. & Moritz, C., Molecular Systematics. Sinauer Ass., Sunderland: 45-126.

Nakamura, H. K. 1986: Chromosomes of Archaeogastropoda (Mollusca: Prosobranchia), with some remarks on their cytotaxonomy and phylogeny. – Publ. Seto Mar. Biol. Lab. 31: 191-267.

Narang, S. K., Kaiser, P. E. & Seawright, J. A. (1989): Identification of species D, a new member of the *Anopheles quadrimaculatus* species complex: a biochemical key. – J. Amer. Mosquito Control Assn. 5: 317-324.

Needleman, S. B. & Wunsch, C. D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. – J. Mol. Biol. 48: 443-453.

Nei, M. (1972): Genetic distance between populations. – Am. Natur. 106: 283-292.

– (1987): Molecular evolutionary genetics. – Columbia University Press, New York.

Nei, M. & Kumar, S. (2000): Molecular evolution and phylogenetics. – Oxford University Press, Oxford.
Nei, M. & Miller, J. C. (1990): A simple method for estimating average number of nucleotide substitutions. – Genetics 125: 873-879.
Nelson, G. (1978): Ontogeny, phylogeny, paleontology, and the biogenetic law. – Syst. Zool. 27: 324-345.
– (1994): Homology and systematics. – In: Hall, B.K. (ed.), Problems of phylogenetic reconstruction. Academic Press, San Diego: 101-149.
Nelson, G. & Platnick, N. (1981): Systematics and biogeography. – Columbia University Press, New York.
Nickrent, D. L. & Starr, E. M. (1994): High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. – J. Mol. Evol. 39: 62-70.
Nikoh, N., Iwabe, N., Kuma, K., Ohno, M., Sugiyama, T., Watanabe, Y., Yasui, K., Shi-cui, Z., Hori, K., Shimura, Y. & Miyata, T. (1997): An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. – J. Mol. Evol. 45: 97-106.
Oeser, E. (1976): Wissenschaft und Information. Systematische Grundlagen einer Theorie der Wissenschaftentwicklung. Bände 1-3, R. – Oldenbourg Verlag, Wien.
– (1987): Das Realitätsproblem. – In: Riedl, R. & Wuketits, F. M. (eds.), Die Evolutionäre Erkenntnistheorie. Verlag Paul Parey, Berlin, Hamburg.
Ohno, S. (1997): The reason as well as the consequences of the Cambrian explosion in animal evolution. – J. Mol. Evol. 44(Suppl.): S23-S27
Ohta, T. (1973): Slightly deleterious mutant substitutions in evolution. – Nature 246: 953-963.
– (1992): The nearly neutral theory of molecular evolution. – Annu. Rev. Ecol. Syst. 23: 263-286.
– (1995): Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. – J. Mol. Evol. 40: 56-63.
– (1997): Role of random genetic drift in the evolution of interactive systems. – J. Mol. Evol. 44 (Suppl.): S9-S14.
Osche, G. (1965): Über latente Potenzen und ihre Rolle im Evolutionsgeschehen. – Zool. Anz. 174: 411-440.
– (1973): Das Homologisieren als eine grundlegende Methode der Phylogenetik. – Aufsätze Reden senckenb. naturf. Ges. 24: 155-165.
– (1985): Wie stehen wir heute zum Biogenetischen Grundgesetz von Haeckel?. – In: Wilhelmi, B. (ed.), "Leben und Evolution". Universität Jena: 56-71.
Osorio, D., Averof, M. & Bacon, J. P. (1995): Arthropod evolution: great brains, beautiful bodies. – TREE 10(11): 449-454.
Owen, R. (1843): Lectures on the comparative anatomy and physiology of the invertebrate animals, delivered at the Royal College of Surgeons, in 1843. – Longman, Brown, Green and Longmans, London.
Page, R. D. M. (1989): Comments on component-compatibility in historical biogeography. – Cladistics 5: 167-182.
Palevody, C. (1969): Donnes sur l'ovogenese d'un Collembole Isotomidae parthenogentique. – C. R. Acad. Sc. Paris 269: 183-186.
Pamilo, P. & Bianchi, N. O. (1993): Evolution of the ZFX and ZFY genes: Rates and independence between the genes. – Mol. Biol. Evol. 29: 180-187.
Panchen, A. L. (1994): Richard Owen and the concept of homology. – In: Hall, B. K. (ed.), "Homology, the hierarchical basis of comparative biology". Academic Press, San Diego: 21-62.
Parker, T. J. (1883): On the structure of the head in *Palinurus*, with special reference to the classification of the genus. – N.Z. J. Sci. 1: 584-585.
Paterson, H. E. A. (1985): The recognition concept of species. – In: Species and speciation (E.S. Vrba, ed.). Transvaal Museum Monograph No. 4. Transvaal Museum, Pretoria.
Patterson, B. & Pascual, R. (1968): Evolution of mammals on southern continents. V. The fossil mammal fauna of South America. – Quart. Rev. Biol. 43: 409-451.
Patterson, C. (1982): Morphological characters and homology. – In: Joysey, K. A. & Friday, A. E. (eds.): Problems of phylogenetic reconstruction. Academic Press, London: 21-74.
– (1988): Homology in classical and molecular biology. – Mol. Biol. Evol. 5: 603-625.
Patterson, C. & Rosen, D. E. (1977): Review of ichtyodectiform and other Mesozoic teleost fishes and the theory and practice of classifying fossils. – Bull. Am. Mus. Nat. Hist. 158: 81-172.
Patton, J. C. & Avise, J. C. (1983): An empirical evaluation of qualitative Hennigian analyses of protein electrophoretic data. – J. Mol. Evol. 19: 244-254.
Pawlowski, J., Bolivar, I., Fahrni, J. F., de Vargas, C., Gouy, M. & Zaninetti, L. (1997): Extreme differences in rates of molecular evolution of Foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record. – Mol. Biol. Evol. 14: 498-505.
Penny, D. & Hendy, M. D. (1985): The use of tree comparison metrics. – Syst. Zool. 34: 75-82.
Perez, M. L., Valverde, J. R., Batuecas, B., Amat, F., Marco, R. & Garesse, R. (1994): Speciation in the *Artemia* genus: mitochondrial DNA analysis of bisexual and parthenogenetic brine shrimps. – J. Mol. Evol. 38: 156-168.
Philippe, H. & Laurent, J. (1998): How good are deep phylogenetic trees? – Curr. Opin. Genet. Dev. 8: 616-623.
Pol, D. & Siddall, M. E. (2001): Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. – Cladistics 17: 266-281.
Polz, H. (1998): *Schweglerella strobli* gen. nov. sp. nov. (Crustacea: Isopoda: Sphaeromatidea), eine Meeres-Assel aus den Solnhofener Plattenkalken. – Archaeopteryx 16: 19-28.

Poore G. C. B. & Lew Ton H. M. (1990): The Hologna-thidae (Crustacea: Isopoda: Valvifera) expanded and redefined on the basis of body-plan. – Invertebr. Taxon. 4:55-80.

Popadic, A., Rusch, D., Peterson, M., Rogers, B. T. & Kaufman, T. C. (1996): Origin of the arthropod mandible. – Nature 380: 395.

Popper, K. R. (1934): Logik der Forschung. – Julius Springer Verlag, Wien. (10. Auflage 1994: J. C. B. Mohr, Tübingen)

Posada, D. & Crandall, K. A. (1998): MODELTEST: testing the model of DNA substitution. – Bioinf. Appl. Note 14: 817-818.

Purvis, A. & Bromham, L. (1997): Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. – J. Mol. Evol. 44: 112-119.

Rambaut, A. & Grassly, N. C. (1997): Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. – Comput Appl Biosci. 13: 235-238.

Rassmann, K. (1997): Evolutionary age of the Galápagos iguanas predates the age of the present Galápagos Islands. – Mol. Phylog. Evol. 7: 158-172.

Rassmann, K., Trillmich, F. & Tautz, D. (1997): Hybridization between the Galápagos land and marine iguana (*Conolophus subcristatus* and *Amblyrhynchus cristatus*) on Plaza Sur. – J Zool Lond 242: 729-739.

Remane, A. (1952): Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik. Theoretische Morphologie und Systematik I. – Akad. Verlagsges. Geest & Portig, Leipzig.

– (1961): Gedanken und Probleme: Homologie und Analogie, Praeadaptation und Parallelität. – Zool. Anz. 166 (9/12): 447-465.

Remane, A., Storch, V. & Welsch, U. (1986): Systematische Zoologie, 3. Auflage. – Gustav Fischer Verlag, Stuttgart.

Richter S. & Meier R. (1994): The development of phylogenetic concepts in Hennig's early theoretical publications (1947-1966). – Syst. Biol. 43: 212-221.

Riedl, R. (1975): Die Ordnung des Lebendigen. – Paul Parey, Hamburg und Berlin.

– (1992): Wahrheit und Wahrscheinlichkeit. – Verlag Paul Parey, Berlin.

Rodrigo, A. G., Kelly-Borges, M., Bergquist, P. R. & Bergquist, P. L. (1983): A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. – N. Z. J. Bot. 31: 257-268.

Rodríguez, F. J., Oliver, A., Marìn, A. & Medina, J. R. (1990): The general stochastic model of nucleotide substitution. – J. Theor. Biol. 142: 485-501.

Roelofs, D. & Bachmann, K. (1997): Coparisons of chloroplast and nuclear phylogeny in the autogamous annual *Microseris douglasii* (Asteraceae: Lactucaceae). – Plant Syst. Evol. 204: 49-63.

Rogers, J. S. (1972): Measures of genetic similarity and genetic distance. – Studies in Genetics VII, Univ. Texas Publ. 7213: 145-153.

Rosa, D. (1918): Ologenesi: Nuova teoria dell'evoluzione e della distribuzione geografica dei viventi. – Bemporad e Figlio, Firenze.

Rzhetsky, A. & Nei, M. (1992): A simple method for estimating and testing minimum-evolution trees. – Mol. Biol. Evol. 9: 945-967.

– (1995): Tests of applicability of several substitution models for DNA sequence data. – Mol. Biol. Evol. 12: 131-151.

Saitou, N. & Nei, M. (1987): The neighbour-joining method: a new method for reconstructing phylogenetic trees. – Mol. Biol. Evol. 4: 406-425.

Saitou, N. (1990): Maximum likelihood methods. – Methods Enzymol. 183: 584-598.

Saller, K. (1959): Der Begriff des Kryptotypus. – Scientia (Bologna) A 53: 158-165.

Schadt, E. E., Sinsheimer, J. S. & Lange, K. (2002): Applications of codon and rate variation models in molecular phylogeny. – Mol. Biol. Evol. 19: 1550-1562.

Schmidt, C. (1999): Phylogenetisches System der Crinochaeta (Crustacea, Isopoda, Oniscidea). – Dissertation, Ruhr-Universität Bochum.

Schminke, H. K. (1987): Le genre *Thermobathynella* Capart, 1951 (Bathynellacea, Malacostraca) et ses relations phylétiques. – Rev. Hydrobiol. Trop. 20: 107-111.

Schneider, T. D. (1996): Information theory primer. ftp.ncifcrf.gov/pub/delila/primer.ps

Scholl A. & Pedroli-Christen A. (1996): The taxa of *Rhymogona* (Diplopoda: Craspedosomatidae): a ring species, part one: genetic analysis of population structure. – Mém. Mus. Natn. Hist. Nat. 169: 45-51.

Scholtz, G. (1995): Head segmentation in Crustacea – an immuncytochemical study. – Zoology 98: 104-114.

– (1997): Cleavage, germ band formation and head segmentation: the ground pattern of the Euarthropoda. – In: Fortey, R. A. & Thomas, R. H. (eds.), Arthropod Relationships. Systematics Association Special Volume Series 55, Chapman & Hall, London, 317-332.

Scholtz, G., Mittmann, B. & Gerberding, M. (1998): The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of Mandibulata. – Int. J. Dev. Biol. 42: 801-810.

Schöniger, M. & Haeseler, A. von (1993): More reliable phylogenies by properly weighted nucleotide substitutions. – In: Opitz, O., Lausen, B. & Klar, R. (eds.), Information and classification; Proc. 16th Ann. Conf. "Gesellschaft für Klassifikation e.V.". Springer Verlag, Berlin: 413-420.

Schram, F. R. (1986): Crustacea. – Oxford University Press, New York & Oxford.

– (1991): Cladistic analysis of metazoan phyla and the placement of fossil problematica. – In: Simonetta A. M. & Morris S. C., (eds.), The early evolution of Metazoa. – Cambridge Univ. Press, Cambridge (England): 35-46.

Schram, F. R. & Emerson, M. J. (1991): Arthropod pattern theory: a new approach to arthropod phylogeny. – Mem. Queensl. Mus. 31: 1-18.
Schubart, C. D., Diesel, R. & Hedges, B. (1998): Rapid evolution to terrestrial life in Jamaican crabs. – Nature 393: 363-365.
Schubart, O. (1934): Tausendfüßler oder Myriapoda. I: Diplopoda. – In: Dahl, F. (Hsg.), Tierwelt Deutschlands 28: 1-318.
Shannon, C. E. (1948): A mathematical theory of communication. – Bell System Tech. J. 27: 379-423, 623-656.
Sharbel, T. F. (1999): Amplified Fragment Length Polymorphisms: a non-random PCR-based technique for multilocus-sampling. – In: Epplen J. T. & Lubjuhn T. (eds.) Methods and Tools in Biosciences and Medicine DNA Profiling and DNA Fingerprinting, Birkhäuser Verlag, Basel: 177-194.
Seibold, I. & Helbig, A. J. (1995): Evolutionary history of New and Old World vultures inferred from nucleotide sequences of the mitochondrial cytochrome b gene. – Philos. Trans. R. Soc. Lond. B 350: 163-178.
Seiffert, H. (1991): Einführung in die Wissenschaftstheorie, Bände 1-3 (9. Auflage). – Beck'sche Verlagsbuchhandlung, München.
Sharp, P. & Li, W. H. (1989): On the rate of DNA sequence evolution in *Drosophila*. – J. Mol. Evol. 28: 398-402.
Sharp, P. M., Tuohy, M. F. & Mosurski, K. R. (1986): Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. – Nucleic Acid Res. 14: 5125-5143.
Shimodaira, H. & Hasegawa, M. (1999): Multiple comparisons of log-likelihoods with applications to phylogenetic inference. – Mol. Biol. Evol. 16: 1114-1116.
Shubin, N., Tabin, C. & Carroll, S. (1997): Fossils, genes and the evolution of animal limbs. – Nature 388: 639-648
Sibley, C. G. & Ahlquist, J. E. (1990): Phylogeny and classification of birds. – Yale Univ. Press, New Haven.
Siddall, M. E. (1998): Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. – Cladistics 14: 209-220.
Siebert, D. J. (1992): Tree statistics; trees and 'confidence'; consensus trees; alternatives to parsimony; character weighting; character conflict and its resolution. – In: Forey, P. L. et al., (eds.) Cladistics – A practical course in systematics. Clarendon Press, Oxford: 72-123.
Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. (1994): Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. – Ann. Entomol. Soc. Am. 87: 651-701.
Simpson, G. G. (1961): Principles of animal taxonomy. – Columbia Univ. Press, New York.
Snodgrass, R. E. (1950): Comparative studies on the jaws of mandibulate arthropods. – Smiths. Misc. Coll. 116: 1-85.
Sober, E. (1986): Parsimony and character weighting. – Cladistics 2: 28-42.
– (1988): Reconstructing the past. Parsimony, evolution, and inference. – The MIT Press, Cambridge (England).
Sokal, R. R. & Rohlf, J. J. (1981): Biometry, second edition. – W. H. Freeman, San Francisco.
Sokal, R. R. & Sneath, P. H. A. (1963): Principles of numerical taxonomy. – W. H. Freeman, San Francisco.
Sorenson, M. D. (1999): TreeRot, version 2 (computer program). – Boston University, Boston, MA.
Spears T., Abele L. G. & Applegate, M. A. (1994): Phylogenetic study of cirripedes and selected relatives (Thecostraca) based on 18S rDNA sequence analysis. – J. Crust. Biol. 14: 641-656.
Starck, D. (1995): Wirbeltiere 5. Teil: Säugetiere. – G. Fischer Verlag, Jena, Stuttgart.
Steel, M. A., Lockhart, P. J. & Penny, D. (1993): Confidence in evolutionary trees from biological sequence data. – Nature 364: 440-442.
Steel, M. A. (1994): Recovering a tree from the leaf colourations it generates under a Markov model. – Appl. Math. Lett. 7: 19-24.
Steel, M. & Penny, D. (2000): Parsimony; likelihood, and the role of models in molecular phylogenetics. – Mol. Biol. Evol. 17: 839-850.
Streble, H. & Krauter, D. (1973): Das Leben im Wassertropfen. – Franckh'sche Verlagsbuchhandlung, Stuttgart.
Strimmer, K. (1997): Maximum likelihood methods in molecular phylogenetics. – Herbert Utz Verlag Wissenschaft, München.
Strimmer, K. & Haeseler, A. von (1996): Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. – Mol. Biol. Evol. 13: 964-969.
Sturmbauer, C. & Meyer, A. (1992): Genetic divergence, speciation and morphological stasis in a lineage of African cichlid fishes. – Nature 358:578-581.
Sudhaus, W. (1980): Problembereiche der Homologienforschung. – Verh. Dtsch. Zool. Ges. 1980: 177-187.
Sudhaus, W. & Rehfeld, K. (1992): Einführung in die Phylogenetik und Systematik. – G. Fischer Verlag, Stuttgart.
Summer, A. T. (1990): Chromosme banding. – Unwin Hyman, London.
Swisher, C. C., Wang, Y. Q., Wang, X. L., Xu, X. & Wang, Y. (1999): Cretaceous age for the feathered dinosaurs of Liaoning, China. – Nature 400: 58-61.
Swofford, D. L. (1990): PAUP: Phylogenetic analysis using parsimony, version 3.0. – Illinois Natural History Survey, Champaign, Illinois.
Swofford, D. L. & Berlocher, S. H. (1987): Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. – Syst. Zool. 36: 293-325.

Swofford, D. L. & Maddison, W. P. (1987): Reconstructing ancestral character states under Wagner parsimony. – Mathem. Biosci. 87: 199-229.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996): Phylogenetic Inference. – In: Hillis, D. M., Moritz, C. & Mable, B. K. (eds.), Molecular Systematics. Sinauer Ass., Sunderland: 407-514.

Szalay, F. S. & Delson, E. (1979): Evolutionary History of the Primates. – Academic Press, New York.

Tajima F. & Nei M. (1984): Estimation of evolutionary distance between nucleotide sequences. – Mol. Biol. Evol. 1: 269-285

Takahaka, N. & Nei, M. (1990): Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. – Genetics 124: 967-978.

Takahashi, K. & Nei, M. (2000): Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. – Mol. Biol. Evol. 17: 1251-1258.

Takezaki, N., Rzhetsky, A. & Nei, M. (1995): Phylogenetic test of the molecular clock and linearized trees. – Mol. Biol. Evol. 12: 823-833.

Tamura, K. (1992): Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. – Mol. Biol. Evol. 9: 678-687.

Tamura, K. & Nei, M. (1993): Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. – Mol. Biol. Evol. 10: 512-526.

Tavaré, S. (1986): Some probabilistic and statistical problems on the analysis of DNA sequences. – Lec. Math. Life Sci. 17: 57-86.

Taylor, W. R. (1986a): Identification of protein sequence homology by consensus template alignment. – J. Mol. Biol. 20: 233-258.

– (1986b): The classification of amino acid conservation. – J. Theor. Biol. 119: 205-218.

Templeton, A. (1983): Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. – Evolution 37: 221-244.

Templeton, A. R. (1989): The meaning of of species and speciation: a genetic perspective. – In: Otte, D. & Endler, J. A. (eds.), Speciation and its consequences. Sinauer, Sunderland: 3-27.

Tengan, C. H. & Moraes, C. T. (1998): Duplictaion and triplication with staggered breakpoints in human mitochondrial DNA. – Biochim. Biophys. Acta 1406: 73-80.

Thenius, E. (1979): Die Evolution der Säugetiere. – UTB/G. Fischer Verlag, Stuttgart.

Thompson J. D., Higgins D. G. & Gibson T. J. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. – Nucleic Acids Res. 22: 4673-4680.

Thorne, J. L., Kishino, H. & Painter, I. S. (1998): Estimating the rate of evolution of the rate of molecular evolution. – Mol. Biol. Evol. 15: 1647-1657.

Thorne, J. L. & Kishino, H. (2002): Divergence time and evolutionary rate estimation with multilocus data. – Syst. Biol. 51: 689-702.

Thorogood, P. (1987): Mechanisms of morphogenetic specification in skull development. – In: Wolff, J. R., Sievers, J. & Berry, M., "Mesenchymal-epithelial interactions in neural development". Springer Verlag, Berlin: 141-152.

Tillier, E. R. M. (1994) Maximum likelihood with multiparameter models of substitution. – J. Mol. Evol. 39: 409-417.

Trontelj, P., Sket, B., Dovc, P. & Steinbrück, G. (1996): Phylogenetic relationships in European erpobdellid leeches (Hirudinea: Erpobdeliidae) inferred from restriction-site data of the 18s ribosomal gene and ITS2 region. – J. Zoo. Syst. Evol. Research 34: 85-93.

Tuffley, C. & Steel, M. (1997): Links between maximum likelihood and maximum parsimony under a simple model of site substitution. – Bull. Math. Biol. 59: 581-607.

Tutt, J. W. (1898): Some considerations of natural genera, and incidental references to the nature of species. – Proc. South Lond. Ent. Nat. Hist. Soc. 1898: 20-30.

Uzzell, T. & Corbin, K. W. (1971): Fitting discrete probability distribution to evolutionary events. – Science 172: 1089-1096.

Van de Peer, Y., Neefs, J. M., De Rijk, P. & De Wachter, R. (1993): Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. – J. Mol. Evol. 37: 221-232.

Van de Peer, Y., Nicolai, S., De Rijk, P. & De Wachter, R. (1996): Database on the structure of small ribosomal subunit RNA. – Nucleic Acids Res 24: 86-91.

Van de Peer, Y. (1997): Variability map of eukaryotic ssu rDNA. http://hgins.uia.ac.be/u/yvdp.

Van Valen, L. (1976): Ecological species, multispecies, and oaks. – Taxon 25: 233-239.

– (1982): Homology and causes. – J. Morphol. 173: 305-312.

Van Raay, T. J. & Crease, T. J. (1995): Mitochondrial DNA diversity in an apomictic *Daphnia* complex from the Canadian High Arctic. – Mol. Ecol. 4: 149-161.

Van Syoc, R. J. (1994): Genetic divergence beween populations of the eastern Pacific goose barnacle *Pollicipes elegans*: mitochondrial cytochrome c subunit 1 nucleotide sequences. – Mol. Mar. Biol. Biotechn. 3: 338-346.

Veron, J. E. N. (1995): Corals in space and time. – UNSW Press, Sydney.

Vollmer, G. (1983): Evolutionäre Erkenntnistheorie. 3. Aufl. – S. Hirzel Verlag, Stuttgart.

Wada, H. & Satoh, N. (1994): Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18SrDNA. – Proc. Natl. Acad. Sci. USA 91: 1801-1804.

Waddell, P. J. (1995): Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. – Ph.D. dissertation, Massey University, New Zealand.

Waddell, P. J. & Steel, M. A. (1997): General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. – Mol. Phylog. Evol. 8: 398-414.

Wägele, J. W. (1982): Isopoda (Crustacea: Peracarida) ohne Oostegite: Über einen Microcerberus aus Florida. – Mitt. Zool. Mus. Univ. Kiel. 1(9): 19-23.

– (1993): Rejection of the "Uniramia" hypothesis and implications of the Mandibulata concept. – Zool. Jb. Syst. 120: 253-288.

– (1987): Description of the postembryonal stages of the Antarctic fish parasite *Gnathia calva* Vanhöffen (Crustacea: Isopoda) and synonymy with *Heterognathia* Amar & Roman. – Polar Biol. 7: 77-92.

– (1989): Evolution und phylogenetisches System der Isopoda: Stand der Forschung und neue Erkenntnisse. – Zoologica 140: 1-262.

– (1994a): Review of methodological problems of 'computer cladistics' exemplified with a case study on isopod phylogeny (Crustacea: Isopoda). – Z. zool. Syst. Evolut.-forsch. 32: 81-107.

– (1994b): Notes on Antarctic and South American Serolidae (Crustacea, Isopoda) with remarks on the phylogeneetic biogeography and a descritpion of new genera. – Zool. Jb. Syst. 121: 3-69.

– (1996): First principles of phylogenetic systematics, a basis for numerical methods used for morphological and molecular characters. – Vie Milieu 46: 125-138.

Wägele, J. W. & Rödding, F. (1998): A priori estimation of phylogenetic information conserved in aligned sequences. – Mol. Phylog. Evol. 9: 358-365.

Wägele, J. W., Erikson, T., Lockhart, P. & Misof, B. (1999): The Ecdysozoa: Artifact or monophylum? – J. Zool. Syst. Evol. Res. 37: 211-223.

Wagner, W. H. (1961): Problems in the classification of ferns. – Rec. Adv. Bot. 1: 841-844.

– (1963): Biosystematics and taxonomic categories in lower vascular plants. – Regnum Veg. 27: 63-71.

Wallis, M. (1997): Function switching as a basis for bursts of rapid change during the evolution of pituitary growth hormone. – J. Mol. Evol. 44: 348-350.

Walossek, D. (1993): The Upper Cambrian *Rehbachiella* and the phylogeny of Branchiopoda and Crustacea. – Fossils Strata 32: 1-202.

Wang, L. S. (2002): Genome Rearrangement Phylogeny Using Weighbor. – Lecture Notes Comp. Sci. 2452: 112-125.

Waterman, M. S. (1984): General methods of sequence comparison. – Bull. Math. Biol. 46: 473-500.

Watrous, L. E. & Wheeler, Q. D. (1981): The out-group comparison method of character analysis. – Syst. Zool. 30: 1-11.

Weiner, J. (1994): Der Schnabel des Finken oder Der kurze Atem der Evolution. – Droemer Knaur, München.

Wenzel, J. W. (2002): Phylogenetic analysis: the basic method. – In: DeSalle, R., Giribet, G. & Wheeler, W.: Techniques in molecular systematics and evolution. Birkhäuser Verlag, Basel: 4-30.

Werman, S. D., Springer, M. S. & Britten, R. J. (1990): Nucleic acids I: DNA-DNA-hybridization. – In: Hillis, D. M. & Moritz, C., Molecular Systematics. Sinauer Ass., Sunderland: 45-126.

Westheide, W. & Rieger, R. (1996): Spezielle Zoologie, Erster Teil: Einzeller und Wirbellose Tiere. – G. Fischer Verlag, Stuttgart.

Wetzel, R. (1995): Zur Visualisierung abstrakter Ähnlichkeitsbeziehungen. – Dissertation, Fakultät für Mathematik der Universität Bielefeld.

Wheeler, Q. D. (1986): Character weighting and cladistic analysis. – Syst Zool 35: 102-109

Wheeler, Q. D. & Meier, R. (eds.) (2000): Species concepts and phylogenetic theory. – Columbia University Press, New York.

Wheeler, W. C. (1990): Combinatorial weights in phylogenetic analysis: a statistical parsimony procedure. – Cladistics 6: 269-275.

– (1996): Optimization alignment: the end of multiple sequence alignment in phylogenetics? – Cladistics 12: 1-9.

– (2000): Heuristic reconstruction of hypothetical-ancestral DNA sequences: sequence alignment vs direct optimization. – In: Scotland, R. & Pennington, R. T. (eds.), "Homology and Systematics". – The Systematics Association Special Volume Series 58, Taylor & Francis, London: 106-113.

– (2002): Optimization alignment: down, up, error, and improvements. – In: DeSalle, R., Giribet, G., Wheeler, W.: Techniques in molecular systematics and evolution. Birkhäuser Verlag, Basel: 55-69.

Wheeler, W. C., Cartwright, P. & Hayashi, C. Y. (1993): Arthropod phylogeny: a combined approach. – Cladistics 9: 1-39.

Wheeler W. C. & Gladstein D. S. (1994): MALIGN: A multiple sequence alignment program. – J. Heredity 85: 417-418.

Wheeler W. C. & Honeycutt R. L. (1988): Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. – Mol. Biol. Evol. 5: 90-96.

Weir, B. S. (1996): Genetic Data Analysis II. Sinauer Ass., Sunderland.

Whiting, M. F., Carpenter, J. C., Wheeler, Q. D. & Wheeler, W. C. (1997): The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. – Syst. Biol. 46: 1-68.

Wiens, J. J. (1995): Polymorphic characters in phylogenetic systematics. – Syst. Biol. 44: 482-500.

Wiley, E. O. (1975): Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. – Syst. Zool. 24: 233-243

– (1978): The evolutionary species concept reconsidered. – Syst Zool 27: 17-26.

– (1980): Is the evolutionary species fiction?-A consideration of classes,individuals and historical entities. – Syst Zool 29: 76-80.
– (1988): Entropy and evolution. – In: Weber B. H., Depew D. J. & Smith J. D. (eds.), Entropy, information, and evolution. Massachusetts Inst. Technol.: 173-188.
Wildman, D. E., Uddin, M., Liu, G., Grossman, L. I. & Goodman, M. (2003): Implications of natural selection in shaping 99.4% nonsynonymous DNA identity beween humans and chimpanzees: enlarging genus *Homo*. – P.N.A.S. 100: 7181-7188.
Wilkerson, R. C., Parsons, T. J., Albright, D. G., Klein, T. A. & Braun, M. J. (1993): Random amplified polymorphic DNA (RAPD) markers readily distinguish cryptic mosquito species (Diptera: Culicidae: *Anopheles*). – Insect Mol. Biol. 1: 205-211.
Williams D. M. (1992): DNA analysis: theory, methods. – In: Forey, P. L., Humphries, C. J., Kitching, I. J., Scotland, R. W., Siebert, D. J. & Williams, D. M. (eds.), Cladistics – a practical course in systematics. Clarendon Press, Oxford: 89-123.
Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A. & Tingey, S. V. (1990): DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. – Nucl. Acids Res. 18: 6531-6535.
Williams, P. L. & Fitch, W. M. (1990): Phylogeny determination using dynamically weighted parsimony method. – In: Doolittle, R. F. (ed.), Methods in enzymology Vol. 183. Academic Press, San Diego: 615-626.
Willmann, R. (1985): Die Art in Raum und Zeit. – Paul Parey, Berlin.
– (1987): Phylognetic systematics, classification and the plesion concept. – Verh. naturwiss. Ver. Hamburg 29: 221-233.
– (1990): Die Bedeutung paläontologischer Daten für die zoologische Systematik. – Verh. Dtsch. Zool. Ges. 83: 277-289.
– (1995): Paläontologie als Evolutionsforschung. Artbildung und Evolutionsfaktoren bei fossilen Organismen. – Veröff. Übersee-Mus. Bremen Naturwiss. 13: 9-30.
Willmer, P. (1990): Arthropod phylogeny. – In: Willmer, P., (ed.) Invertebrate relationships – patterns in animal evolution. – Cambridge Univ. Press, Cambridge: 271-299.
Wills, M. A., Briggs, D. E. G., Fortey, R. A. & Wilkinson, M. (1995): The significance of fossils in understanding arthropod evolution. – Verh. Dtsch. Zool. Ges. 88.2: 203-215.
Wink, M. (1995): Phylogeny of Old and New World vultures (Aves: Accipitridae and Cathartidae) inferred from nucleotide sequences of the mitochondrial cytochrome b gene. – Z. Naturforsch. 50c: 868-882.
Wistow, G. (1993): Identification of lens crystallin: a model system for gene recruitment. – Methods Enzymol. 224: 563-575.
Wolters, J. (1991): The troublesome parasites – molecular and morphological evidence that the Apicoplexa belong to the dinoflagellate-ciliate clade. – BioSystems 25: 75-83.
Woodburne, M. O. & Zinsmeister, W. J. (1984): The first land mammal from Antarctica and its biogeographic implications. – J. Paleont. 58: 913-948.
Wu, C. I. & Li, W. H. (1985): Evidence for higher rates of nucleotide substitution in rodents than in man. – Proc Natl Acad Sci USA 82: 1741-1745.
Yang, Z. (1994): Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximative methods. – J. Mol. Evol. 39: 306-314.
– (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. – Syst. Biol. 43: 329-342.
– (1996): Among-site rate variation and its impact on phylogenetic analyses. – TREE 11: 367-372.
Yang, Z. & Yoder, A. D. (2003): Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and clibration points, with application to a radiation of cute-looking mouse lemur species. – Syst. Biol. 52: 1-12.
Yeates, D. K. (1995): Groundplans and exemplars: paths to the tree of life. – Cladistics 11: 343-357.
Yoder, A. D. & Yang, Z. (2000): Estimation of primate speciation dates using local molecular clocks. – Mol. Biol. Evol. 17: 1081-1090.
Young, N. D. & Healy, J. (2003): GapCoder automates the use of indel characters in phylogenetic analysis. – BMC Bioinformatics 4: 6.
Zabeau, M. & Vos, P. (1993): Selective restriction fragment amplification: a general method for DNA fingerprinting. – European Patent Office publ. 0 534 858 A1.
Zardoya, R. & Meyer, A. (1996): The complete nucleotide sequence of the mitochondrial genome of the lungfish *(Protopterus dolloi)* supports its phylogenetic position as a close relative of land vertebrates. – Genetics 142: 1249-1263.
Zhang, J. & Nei, M. (1996): Evolution of *Antennapedia*-class homeobox genes. – Genetics 142: 295-303.
Zhang, J. (2000): Rates of conservative and radical nonsynobymous nucleotide substitutions in mammalian nuclear genes. – J. Mol. Evol. 50: 56-68.
Zharkikh, A. (1994): Estimation of evolutionary distances between nucleotide sequences. – J. Mol. Evol. 39: 315-329.
Zischler, H., Geisert, H. & Castresana, J. (1998): A hominoid-specific nuclear insertion of the mitochondrial D-loop: implications for reconstructing ancestral mitochondrial sequences. – Mol. Biol. Evol. 15: 463-469.

17. Index

D

E

F

G

H

I

O

P

Q

R

S

T

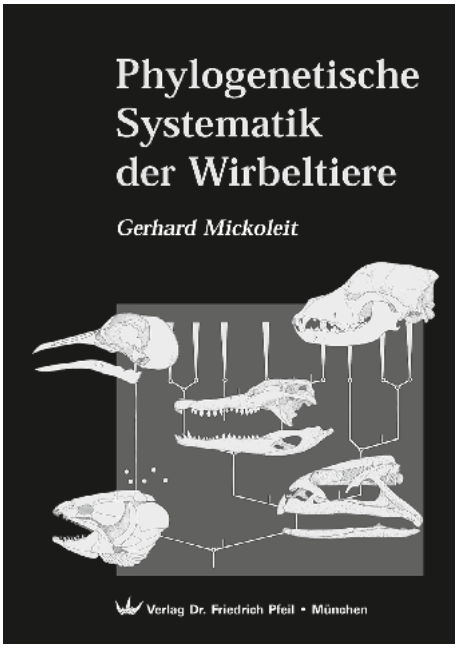
U

V

W

X

Y



MICKOLEIT, Gerhard:

Phylogenetische Systematik der Wirbeltiere

2004. – 675 pp., 676 b/w figures

24.5 × 17.3 cm. Hard cover

ISBN 3-89937-044-9

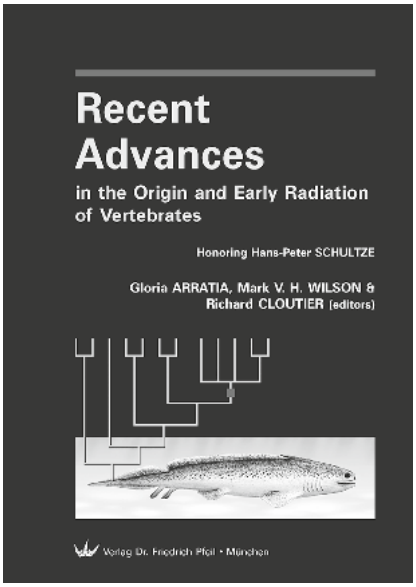
Euro 98.00

Die »Phylogenetische Systematik der Wirbeltiere« des Tübinger Zoologen Gerhard Mickoleit sollte ursprünglich als dritter und abschließender Teil des von Willi Hennig begründeten »Taschenbuchs der Speziellen Zoologie« erscheinen. Durch unermüdliche jahrzehntelange Arbeit gelang es dem Autor aber, weit über den Rahmen des »Taschenbuchs« hinauszugelangen. So liegt nun ein Kompendium vor, in dem die phylogenetischen Verwandtschaftsbeziehungen der rezenten Wirbeltier-Teilgruppen umfassend dargestellt und begründet werden. Darüber hinaus gewährt das Buch Einblicke in den evolutiven Wandel der Organsysteme. Die Darstellung folgt methodisch der von Willi Hennig entwickelten, heute allgemein anerkannten Konsequent-Phylogenetischen Systematik. Im Vordergrund der Darstellung stehen deshalb die Apomorphien, die die monophyletischen Gruppen begründen und die Verwandtschaftshypothesen stützen. Sie konzentrieren sich weitgehend auf den Bereich der Strukturforschung, insbesondere auf die makroskopische Anatomie und die Embryologie. Ausführlich behandelt werden die höheren (im Paläo- und Mesozoikum entstandenen) supraspezifischen Kategorien des Systems. Familien werden in kurzen Diagnosen vorgestellt. Zu jeder behandelten Familie werden exemplarisch einige Arten genannt, die auf Grund anatomischer, ethologischer oder anderer Besonderheiten Bedeutung erlangt haben. Mit ihrer Aufnahme wird dem Benutzer des Buches die Möglichkeit gegeben, sich bei den in der wissenschaftlichen Vertebraten-Literatur häufig genannten Arten hinsichtlich deren Stellung im System und in der Hierarchie der Grundpläne einen Überblick zu verschaffen.

Die besprochenen Merkmale werden mit 676 Abbildungen, größtenteils speziell für dieses Buch angefertigt, veranschaulicht. Die phylogenetischen Zusammenhänge werden mit schematischen Stammbäumen dargestellt. Ausführliche Register für Tiernamen und Fachbegriffe ermöglichen dem Leser gezieltes Nachschlagen. Das Buch richtet sich an die Studierenden und Lehrenden der Zoologie, aber auch an den Wirbeltierspezialisten, der sich außerhalb seines engeren Arbeitsgebietes kundig machen möchte.

Der Autor ist Akademischer Oberrat a.D. am Lehrstuhl für Spezielle Zoologie und war Leiter der Sammlungen des Zoologischen Institutes der Universität Tübingen.

Verlag Dr. Friedrich Pfeil
Wolfratshauser Str. 27, D–81379 München, Germany
phone: + 49-(0)89-74 28 270 – fax: + 49-(0)89-72 42 772 – e-mail: info@pfeil-verlag.de
www.pfeil-verlag.de



ARRATIA, Gloria, Mark V. H. WILSON & Richard CLOUTIER (editors):

Recent Advances

in the Origin and Early Radiation of Vertebrates

Honoring Hans-Peter Schultze

2004. – 703 pp., 5 color and 264 b/w figures,
23 tables, 17 appendices

24.5 × 17.3 cm. Hard cover

ISBN 3-89937-052-X

Euro 240.00

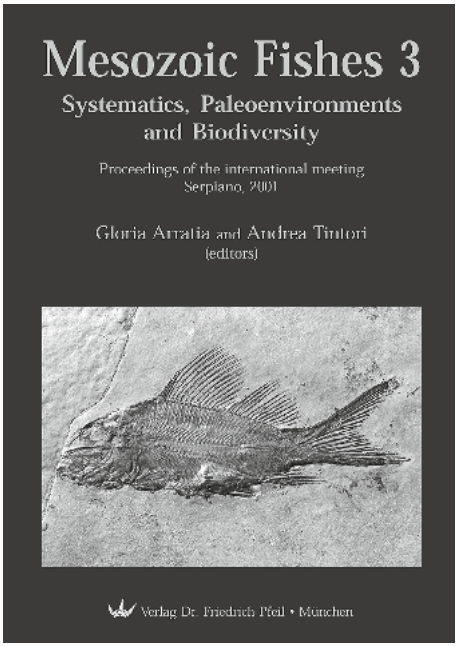
The first discoveries of Early Paleozoic fishes took place in Scotland and in the Baltic area at the beginning of the 19th century. The first early vertebrate remains recorded from Scotland were of Carboniferous age and are now referred to the sarcopterygians *Rhizodus* and *Megalichthys*. Later, discoveries of additional Scottish and Baltic localities made these regions (and also European workers) the main source of information on early vertebrates for a long time. This situation reached its most important development with the contributions of E. STENSIÖ and other Swedish and Danish colleagues, who organised important collecting expeditions (e.g., Podolia and Spitsbergen). New material from these localities and others (e.g., Devonian localities of eastern Canada) allowed STENSIÖ and his followers (the so-called Swedish School) to produce some fascinating morphological work and to propose hypotheses about the origin of early tetrapods that still today are a source of discussion.

New scientific findings have the potential to produce considerable changes in previous interpretations. Vertebrates are not an exception. Based on information gathered over almost two centuries it has long been believed that the origin of vertebrates occurred "*sometime*" during the earliest Paleozoic, "*somewhere*" in the northern Hemisphere. However, discoveries of early vertebrates in the Southern Hemisphere (e.g., Australia and Bolivia) led to a new understanding of the early history of the group. These new discoveries have been remarkable in stimulating new collecting. Recent progress has included the discovery of the "earliest" forms in the Lower Cambrian of China together with new and controversial interpretations of the conodonts.

The most recent decade saw new findings that concern not only the earliest vertebrates, but also most fish groups as well as lower tetrapods. They shed new light on the origin and diversification of basal vertebrates and gnathostomes. Critical fossils have been discovered in many different parts of the world. This new material is having a significant impact on previous character interpretation and distribution, as well as on previous phylogenetic hypotheses.

This book brings together many of these recent discoveries and new interpretations to commemorate the retirement of Hans-Peter SCHULTZE from the Museum für Naturkunde in Berlin. H.-P. SCHULTZE has worked on most groups of lower vertebrates ranging from conodonts to early tetrapods. He has collected in most of the crucial sites around the world. He is one of the most productive researchers in paleoichthyology and is considered by many to be the leading figure in this field.

Verlag Dr. Friedrich Pfeil
Wolfratshauser Str. 27, D–81379 München, Germany
phone: + 49-(0)89-74 28 270 – fax: + 49-(0)89-72 42 772 – e-mail: info@pfeil-verlag.de
www.pfeil-verlag.de



ARRATIA, Gloria & Andrea TINTORI (editors):

Mesozoic Fishes 3

Systematics, Paleoenvironments and Biodiversity

2004. – 649 pp., 19 color and 277 b/w figures, 25 tables, 19 appendices

24.5 × 17.3 cm. Hard cover

ISBN 3-89937-053-8

Euro 240.00

The Mesozoic was an important time in the evolution of chondrichthyan and actinopterygian fishes because it was then that most of the modern groups first entered the fossil record and began to radiate. By the end of the era, many archaic forms had disappeared and the foundation had been laid for the ichthyofauna that now exists. Despite this significant evolutionary change, before 1990 there had been little concerted research done on Mesozoic fishes and no synopsis or compilation of the systematics and paleoecology of Mesozoic fishes had been published, not even for single groups. To remedy this deficiency, Gloria ARRATIA initiated the symposium "Mesozoic Fishes". The first meeting "Mesozoic Fishes – Systematics and Paleoecology" was held in Eichstätt from August 9 to 12, 1993 and the first volume of Mesozoic Fishes, including 36 papers concerning elasmobranchs, actinopteygians and sarcopterygians and the paleoecology of certain important fossil localities was published in 1996. Gloria ARRATIA and Hans-Peter SCHULTZE organized the second Symposium. It was held in Buckow, from July 6 to 10, 1997. The results of the symposium were published in "Mesozoic Fishes 2 – Systematics and Fossil Record" and included 31 papers.

Andrea TINTORI, Markus FELBER and Heinz FURRER organized the third Symposium. It was held in Serpiano, Monte San Giorgio from August 26 to 31, 2001.

The results of the symposium presented in this volume reflect the current state of knowledge of Mesozoic fishes. Evaluation of major fish groups such as Mesozoic chondrichthyans, halecostomes and sarcopterygians and of the Mesozoic fossil record of continents such as North America, Asia, South America and Africa are the central issue. In addition, new information on chondrichthyans, actinopterygians and sarcopterygians are presented. The new findings and the evaluations of the present state of knowledge of Mesozoic fishes described in 33 papers are an exciting invitation to further research.

Verlag Dr. Friedrich Pfeil
Wolfratshauser Str. 27, D–81379 München, Germany
phone: + 49-(0)89-74 28 270 – fax: + 49-(0)89-72 42 772 – e-mail: info@pfeil-verlag.de
www.pfeil-verlag.de

Phylogeny inference and the classification of organisms are indispensable for all fields of biology. On the basis of a well corroborated tree of life it is possible to understand the evolution of structure and function, of genomes, of gene families, of cascades of developmental genes, and the origin of genes of medical importance. Ecologists need a stable classification of organisms to identify organisms, to find their correct names and thus further information on relevant species. This book offers an introduction to the theory of Phylogenetic Systematics and is a companion for all biologists who want to analyze morphological or molecular data with classical methods or with modern computer programs. The first part of the book explains the epistemological basis that is independent of the type of method used to construct phylogenetic trees. Unlike other empirical sciences, the estimation of data quality in phylogenetics is still little developed and very often neglected. Here a theoretical basis is presented that enables the systematist to assess critically and objectively the quality of different data sets and to make statements on the plausibility of results. This requires a conception of the notions of *information content*, *probability of homology*, *probability of cognition*, *probability of events*, the *principle of parsimony*, the differentiation of *phenomenological* and *modelling* methods. Willi Hennig's original method is compared with modern numerical systematics and an updated Hennigian procedure of data analysis is discussed. The difference between phenetic and phylogenetic cladistics is explained. Popular tools for data evaluation implemented in computer programs are explained including their axiomatic assumptions, sources of error and possible applications. For the more common tools the mathematical background is explained in a simple, easy-to-understand way.

Johann-Wolfgang Wägele was until recently head of the Department for Animal Systematics (Lehrstuhl für Spezielle Zoologie) at the University of Bochum and is now director of the Museum Alexander Koenig in Bonn (Germany). His main research interests are the taxonomy, phylogeny and biodiversity of Isopoda, which implies observations of life history, biogeography and ecology in combination with phylogeny inference. Further subjects include arthropod phylogeny and tools for explorative data analyses. The author is president of the Gesellschaft für Biologische Systematik, a Central European society of systematists, and he is actively promoting biodiversity research.

ISBN 3-89937-056-2