| SPIXIANA | 37 | 1 | 21–26 | München, August 2014 | ISSN 0341–8391 |
|---|---|---|---|---|---|

# How to use CAOS software for taxonomy? A quick guide to extract diagnostic nucleotides or amino acids for species descriptions

**Katharina M. Jörger & Michael Schrödl**

Jörger, K. M. & Schrödl, M. 2014. How to use CAOS software for taxonomy? A quick guide to extract diagnostic nucleotides or amino acids for species descriptions. Spixiana 37 (1): 21–26.

The inclusion of molecular characters into species descriptions is becoming increasingly accepted in the taxonomic community. Morphologically cryptic species might even require molecular taxonomy to be delimited and diagnosed. In current absence of established standard procedures and software, the practical application of molecular taxonomy proved to be non-trivial and fraught with potential pitfalls. Here we present a step-by-step guide how to extract diagnostic molecular characters via the Character Attribute Organization System (CAOS) software from sequence alignments to be used in formal species descriptions. We highlight the necessary technical and general considerations when extracting diagnostic characters from molecular sequence data and argue for using them within integrative taxonomic frameworks.

Katharina M. Jörger (corresponding author) & Michael Schrödl, SNSB – Bavarian State Collection of Zoology, Münchhausenstr. 21, 81247 München, Germany; and Department Biology II, BioZentrum, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany;
e-mail: joerger@bio.lmu.de, Michael.Schroedl@zsm.mwn.de

## Introduction

In the past years, molecular characters have emerged in taxonomic descriptions in various, inconsistent forms, mainly as pure additives to morphology-based diagnoses, but seldom making use of the diagnostic content of the molecular data itself (Goldstein & DeSalle 2011). In absence of an established standard on molecular taxonomy, Jörger & Schrödl (2013) discussed the theoretical background and practical considerations when founding species descriptions (solely) on molecular characters. The authors practically applied molecular taxonomy by describing nine species of cryptic acochlidian sea slugs (Gastropoda: Panpulmonata) based on molecular characters, i.e. diagnostic nucleotides from four molecular markers, both nuclear and mitochondrial. These species had been previously discovered in a combined molecular species delineation approach (Jörger et al. 2012). Jörger & Schrödl (2013) used the Character Attribute Organization System (CAOS) software (Sarkar et al. 2002, Sarkar et al. 2008, Bergmann et al. 2009) to retrieve the diagnostic information, which required manual steps to adapt it for their purposes. Here, we briefly provide a preliminary step-by-step guide of using CAOS for molecular taxonomy. This aims to support other taxonomists until the available software is modified for better suiting the needs of the taxonomic community. We also comment on what we think could be minimum requirements for using our protocol for describing new species.

## Protocol using CAOS

### Before getting started

**Selection of an appropriate evolutionary entity:** Based on a given topology, CAOS can serve to extract diagnostic nucleotides which enable to distinguish a clade from its sister clade at a given node of a tree. If a species is only compared to its supposed direct sister species, 1) the selected diagnostic characters depend entirely on the validity of this phylogenetic sister group relationship, and 2) it increases the risk of including plesiomorphies as diagnostic characters; CAOS diagnoses, e. g., species A with 'G' at position 100 of the alignment, if species B bears another nucleotide, even if all other included lineages equally present 'G' at position 100. It is therefore advisable to increase the range across which diagnostic characters are determined, e. g. comparing species A not only to its more or less insecurely inferred or supposed sister species, but, e. g. to all available congeners (as done by Jörger & Schrödl 2013) or all members of any suitable, i. e. well-supported and more inclusive clade. Given only four potential possibilities in a homologous position of a nucleotide alignment there is a risk that multiple mutations create homoplastic character states and this risk increases with evolutionary distance (Rach et al. 2008). Therefore (and for their potential of lowering the alignment quality) distant outgroups should not be included when using CAOS analyses for species descriptions, especially when dealing with fast evolving markers. The selection of the evolutionary entity, which serves for comparison is probably the most fundamental decision in the analyses, i. e. choose the best supported monophylum, which comprises the radiation of closely related species, including the one of taxonomic interest.

**Relevance of the input tree in CAOS analyses:** The input tree serves as mere topological guide for the algorithm to determine which uploaded sequences are compared with each other. Therefore, it can be retrieved by running a quick phylogenetic analysis, e. g. using Geneious Tree Builder in Geneious (Biomatters, http://www.geneious.com/); RAxML (Stamatakis 2006) software, or might as well be written by hand for smaller datasets. The input tree for CAOS can be manipulated in order to determine the diagnostic characters of species A in relation to any others (not necessarily reflecting best possible phylogenetic hypotheses). You can define species A as single terminal of a clade that is sister to a clade uniting all its congeners, for example (see Procedure in CAOS below).

**Diagnostic characters in molecular taxonomy:** CAOS offers the possibility to extract different character attributes (CAs) (Sarkar et al. 2002, Rach et al. 2008, Sarkar et al. 2008, Bergmann et al. 2009). For molecular taxonomy, Jörger & Schrödl (2013) consider only single pure CAs (sPu) relevant, i. e. characteristic nucleotides (or amino acids) present in members of species A but absent from all members of its sister clade at a given node. The program further distinguishes homogeneous sPu (i.e. present in all specimens of species A) and heterogeneous sPu (i. e. different characters states in species A which are, however, absent from the species under comparison). The latter might be problematic due to potentially convergently evolved character states (Jörger & Schrödl 2013) and should therefore only be used as additional information to the diagnostic homogeneous sPu's in a species description.

**Positional homology assumptions:** The alignment presents the positional homology assumptions of the dataset and its quality is crucial for reliability of molecular based taxonomy. Jörger & Schrödl (2013) demonstrated how quality and quantity of the diagnostic nucleotides might vary with different algorithms applied and how errors in the alignment can artificially inflate detected diagnostic characters. Therefore, it is indispensable to critically compare the performance of different alignment programs on your dataset, e. g. using Muscle (Edgar 2004) and Mafft (Katoh et al. 2002, Katoh et al. 2005); if regions of the alignment are ambiguously aligned it may be useful to mask the alignment, e. g. with Gblocks (Talavera & Castresana 2007) or Aliscore (Misof & Misof 2009); this reduces the number of diagnostic characters but increases their reliability. After comparing different approaches rely on the most conservative one to avoid to artificially inflating your diagnoses. For the reproducibility of taxonomic characters all steps within the alignment procedure need to be reported in detail and manual changes in the alignment are not acceptable unless they are objectively justified and appropriately documented, to be reproducible in future research.

**Traceability and testability of molecular species diagnoses:** To allow for a maximum in traceability, ideally sequences should be used for the alignment exactly as submitted to or as retrieved from public databases (e. g. GenBank, http://www.ncbi.nlm.nih.gov/genbank/). If you truncated or edited them in some sort, this needs to be clearly stated in the publication and must be reproducible in future research (e. g. two bases trimmed differently in the beginning of the sequence will lead to completely wrong determination of nucleotide positions when someone

intents to reproduce the data). The alignments used for analyses should be either directly be added to the species description (e. g. as supplementary material) or deposited in a public repository such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/).

**Procedure in CAOS**

It is recommended to run the CAOS analyses for each molecular marker separately. If you use concatenated alignments, you need to recalculate the corresponding alignment positions of each diagnostic nucleotide. In case you have identical sequence names and counts on the different markers, you can reuse your input tree (= topology) for each run.

Getting started: for CAOS analysis you need a non-interleaved Nexus-format of your alignment and a corresponding tree (i. e. with identically named sequences). The tree needs to be fully resolved and must not contain polytomies. The CAOS manual provides information how to generate this matrix+tree Nexus file via Mesquite (Maddison & Maddison 2011) or MacClade (Maddison & Maddison 2005). We noted three important steps for the preparation of matrix+tree Nexus file via Mesquite, which are not included in CAOS manual: after open your alignment Nexus file in Mesquite you have to incorporate your tree file (see http://boli.uvm.edu/caos-workbench/manual5.php). At this point, you may have duplicate, identical matrices. One of them should be deleted. Also, if you manipulate manually the tree (to resolve polytomies or to rearrange clades) in Mesquite, remember to 'store tree' (go to Tree section in the Mesquite heading menu) before saving your matrix+tree Nexus file. Last, after saving your file, you may need to open it on a text editor (e. g. NotePad ++, http://notepad-plus-plus.org/) and manually remove the word "matrix" from the text (in general, situated in line 18)[1].

You will need your identical single marker alignment ready in Fasta-format to the next step in CAOS (CAOS-Barcoder), to map the CAs. Pay attention that the order of taxon names is identical in your tree block and in the character matrix, this will otherwise hinder CAOS analyses.

Open CAOS-workbench: http://boli.uvm.edu/caos-workbench/caos.php and the CAOS Analyser: upload your Nexus file and perform CAOS analyses. Save 'CAOS-attribute' and 'CAOS-group'-file. Go to CAOS Barcoder: upload your attribute and group file and your alignment in Fasta-format. Choose whether it presents a nucleotide or amino acid alignment and select what character attributes (CA) you wish the program to extract. There are six options which can be selected, each one resulting in an output Excel-Table numbered according to the selected option. For single pure CAs, for example, check "All sPu characters (homogeneous and heterogeneous)", which is the third option to be selected. Accordingly, your output table will be named "overview3". This table is your result summary, it provides you with all single pure CAs for each lineages at a given node of the input tree in relation to the sister lineage. Ideally, you placed your species (or lineage) of interest in the basal position at a node where it opposes all remaining lineages.[2]

If the diagnostic characters for one clade were all you needed, you just have to repeat the analyses with your alignments generated with different algorithms (in case they diverge) and critically compare the results with regards to the amount and identity of diagnostic characters. Then you are done. If you are interested in the diagnostic characters of several (or all) lineages in your dataset, the basic procedure now needs to iteratively modified and repeated depending on your taxonomic needs. If you now want to proceed extracting the diagnostic characters of species B in relation to all other included sequences, you need to manually reroot the input tree in Mesquite placing the lineage of interest sister to all remaining sequences (e. g. in Mesquite using the 'reroot at branch tool' in the tree window), then store the tree and save the file (e. g. under CAOS-input file_species B).

**Evaluation and presentation of results**

It is crucial for molecular based or supported species descriptions that the results are traceable and reproducible in future research:

Jörger & Schrödl (2013) already underlined the need to deposit the alignments to a public database or supplement it as additional material to the publication of the species description to have it accessible to future research.

The resulting output Excel-Table in CAOS lists the position of the diagnostic CAs within the alignment. Report these positions in your species diagnoses, but additionally select a reference se-

1 If you do not delete the word "matrix" from the Nexus file using a text editor a bug may occur which causes empty files in CAOS or the CAOS-server to crash.

2 CAOS seems to have a problem with missing data. It might not reliably discriminate between gaps and '?' or 'N'. If you have 'N's (e. g. because one sequence at the node is shorter than the others), CAOS sometimes fails to recognise gaps or even true CAs as such. Therefore, in this case results at the problematic positions need to be checked manually.

quence of the species and report the position of the CAs therein. This can be identical to the alignment position but might vary in case of insertions present in other aligned sequences. By adding new data in future research, the alignment might deviate from the one you presented and a reference sequence helps to trace the character positions of interest. Ideally, reference sequences should be generated from type material, or conspecifity needs to be justified by other means. Technically, the positions within a reference sequences can be easily retrieved when viewing the original (unmasked) alignment in programs of DNA sequence analyses such as e. g. Geneious (select Properties: show original base numbers[3]).

## Conclusions

The described procedure overcomes the burden of extracting diagnostic nucleotides from alignments by eye, but is still time-consuming if executed properly. As noted, our protocol requires several indispensable steps before getting started with using CAOS software, such as evaluating various alignments, and iterative application of CAOS requires several manual adjustments. This protocol was applied by Jörger & Schrödl (2013) as a showcase for using diagnostic nucleotide characters for formal descriptions of new species. It is important to note that these new species were previously delineated in an integrative framework and using combined evidence from various molecular approaches (Jörger et al. 2012). In addition, the recovery of numerous, supposedly fixed diagnostic nucleotides for certain lineages in both mitochondrial and nuclear markers (Jörger & Schrödl 2013) can be seen as supporting evidence for their species status under diagnosable phylogenetic and unified species concepts (De Queiroz 2007). Maintenance of diagnostic nucleotides in sympatric or even syntopic populations would also indicate reproductive isolation, i. e. biological species.

In our protocol using CAOS, the sPu's found for a certain species entity are diagnostic as compared to other samples included into analysis, but not all sPu's are necessarily synapomorphic. If it is desired to increase the ratio of putatively apomorphic to plesiomorphic nucleotides among sPu's, the outgroup taxon sampling needs to be expanded, i. e. beyond including members of the target organism's next most inclusive reliable clade (e. g. genus). However, this will increase the risk of undetected multiple substitutions, diminish the number of sPu's for the targets, and thus reduce the diagnostic power. For their higher probability of homology we prefer using homogeneous rather than heterogeneous sPu's for molecular taxonomy, but recommend adding such information into formal species descriptions. Heterogeneous sPu's can be informative as well, e. g. when working at an infraspecific level, comparing divergence and evolution of allopatric populations. Other (e. g. private or any compound) CA's extracted by CAOS may even more sensibly diagnose early stages of genetic isolation between populations and thus may be used within descriptions of otherwise reliably delineated species. For conservativeness, we hesitate to make use of them for formal descriptions of morphologically cryptic species.

Our protocol applied to single sequence markers (nucleotides or amino acids) can provide additional support for species entities, and is herein promoted to be used for formal description of already discovered molecular lineages, e. g., morphologically problematic or cryptic species. However, this should be always done within an integrative taxonomic context (see Jörger et al. 2012, Jörger & Schrödl 2013). If new species are largely or exclusively based on molecular evidence, then more than a single gene marker should be used to provide more information and stability in species descriptions (Jörger & Schrödl 2013). Moreover, this claim addresses the well-known fact that gene trees are not necessarily identical with species trees, e. g. because of incomplete lineage sorting or horizontal gene transfer. We acknowledge that mitochondrial markers more rapidly reflect recent speciation than most nuclear markers (e. g. Birky 2013), but they have specific in their natural history that may cause problems in species delineation and diagnosis (e. g. introgression, see Ballard & Whitlock 2004). We thus propose that for formal species descriptions based on diagnostic sequences, mitochondrial markers should always be supplemented by at least one informative nuclear marker. As a bad example, this CAOS-based or a similar protocol could be used for formally establishing new species from global genetic databases, e. g. extracting sPu's from the barcoding region of the mitochondrial COI gene and naming dozens or hundreds of so far unnamed BINs (Ratnasingham & Hebert 2013) from BOLD. We emphasize that any taxonomic acts should be done by taxonomists and only within a comprehensive revisory context including other genetic and phenotypic data and methods. As with morphology-based taxonomy, there is always room for misuse; molecular taxonomy is a young discipline and with these initial recommendations we want to help to establish a good,

3 Attention: If you have inserted 'N's or '?' in the beginning of a sequence due to different sequence length, these need to be removed since they are otherwise counted, but will not appear once you deposit the sequence to GenBank!

sustainable practice. Molecular taxonomy can be and should be embedded within traditional taxonomy.

## Outlook

To compensate for phylogenetic insecurity (i. e. the probabilistic nature of sister clade hypotheses) and for potential inadequacy of taxon and data sampling, it is crucial that potential species entities are compared with all relevant members of a reliable, more inclusive clade, as performed semi-manually and iteratively by Jörger & Schrödl (2013), with technical and procedural details given herein. Alternatively, software such as SPIDER (Brown et al. 2012) based on the statistical programming environment R can be used to retrieve diagnostic nucleotides from predefined species units. Currently, the "nugDiag" function in SPIDER is limited to extracting single pure diagnostic characters sensu Sarkar et al (2008) (see manual http://spider.r-forge.r-project.org/docs/spider-manual.pdf and Kekkonen & Hebert (2014)). Applying CAOS thus is more tedious but also more flexible, and is suitable to retrieve signal from recent speciation events and for tracing intraspecific nucleotide evolution. Alignment testing and careful selection and variation of ingroups for comparison as described herein are essential for all approaches on diagnostic nucleotides.

We believe that implementing such considerations into future, automated versions of CAOS (or independent attempts such as SPIDER) would create a powerful and efficient tool for diagnosing and describing new species. Within an integrative taxonomic framework, automated procedures based partly or entirely on diagnostic nucleotides would facilitate or at least greatly speed up delimitations and descriptions of morphologically problematic or even fully cryptic species. Hopefully, our CAOS-based, preliminary procedure already may encourage integrative taxonomists to proceed to the final step, i. e. not only delineating but also describing and naming new cryptic species, making them available for biodiversity research.

## Acknowledgements

We wish to thank Tjard Bergmann (Tierärztliche Hochschule, Hannover) for his support with CAOS software and Martin Spies (ZSM) for valuable discussion on the principles of taxonomy and nomenclature. Thanks also to Vinicius Padula (ZSM) for critically testing the workflow and his help in improving the guidelines.

## References

Ballard, J. W. O. & Whitlock, M. C. 2004. The incomplete natural history of mitochondria. Molecular Ecology 13: 729–744.

Bergmann, T., Hadrys, H., Breves, G. & Schierwater, B. 2009. Character-based DNA barcoding: a superior tool for species classification. Berliner und Münchener Tierärztliche Wochenschrift 122: 446–450.

Birky, C. W. 2013. Species detection and identification in sexual organisms using population genetic theory and DNA sequences. PLoS ONE 8: e52544.

Brown, S. D., Collins, R. A., Boyer, S., Lefort, M. C., Malumbres-Olarte, J., Vink, C. J. & Cruickshank, R. H. 2012. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources 12: 562–565.

De Queiroz, K. 2007. Species concepts and species delimitation. Systematic Biology 56: 879–886.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

Goldstein, P. Z. & DeSalle, R. 2011. Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. Bioessays 33: 135–147.

Jörger, K. M. & Schrödl, M. 2013. How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology 10: 59.

– – , Norenburg, J. L., Wilson, N. G. & Schrödl, M. 2012. Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. BMC Evolutionary Biology 12: 245.

Katoh, K., Kuma, K., Toh, H. & Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33: 511–518.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059–3066.

Kekkonen, M. & Hebert, P. D. N. 2014. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Molecular Ecology Resources 14 (4): 706-715. Doi:10.1111/1755-0998.12233.

Maddison, D. R. & Maddison, W. P. 2005. MacClade 4: analysis of phylogeny and character evolution. Version 4.08.

Maddison, W. P. & Maddison, D. R. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Misof, B. & Misof, K. 2009. A monte carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Systematic Biology 58: 21–34.

Rach, J., DeSalle, R., Sarkar, I. N., Schierwater, B. & Hadrys, H. 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proceedings of the Royal Society B, Biological Sciences 275: 237–247.

Ratnasingham, S. & Hebert, P. D. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) System. PLoS One 8: e66213.

Sarkar, I. N., Planet, P. J. & DeSalle, R. 2008. CAOS software for use in character-based DNA barcoding. Molecular Ecology Resources 8: 1256–1259.

–– , Thornton, J. W., Planet, P. J., Figurski, D. H., Schierwater, B. & DeSalle, R. 2002. An automated phylogenetic key for classifying homeoboxes. Molecular Phylogenetics and Evolution 24: 388–399.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

Talavera, G. & Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56: 564–577.