CAOS Software for Use in Character Based DNA Barcoding

Indra Neil Sarkar[1,*], Paul J. Planet[2] and Rob DeSalle[2]

1) MBLWHOI Library, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543 USA
2) Sackler Institute for Comparative Genomics, American Museum of Natural History, 79th Street at CPW, New York, NY 10024

Keywords: Barcoding, Character Analysis, Bioinformatics, Decision Support

*To Whom Correspondence Should Be Sent:
Indra Neil Sarkar, PhD
MBLWHOI Library
Marine Biological Laboratory
7 MBL Street
Woods Hole, MA 02543 USA
Fax: +1 508 540 6902
Email: sarkar@mbl.edu

Running Title: CAOS Software for DNA Barcoding

Abstract
The success of character based DNA barcoding depends on the efficient identification of diagnostic character states from molecular sequences that have been organized hierarchically (e.g., according to phylogenetic methods). Similarly, the reliability of these identified diagnostic character states must be assessed according to their ability to diagnose new sequences. Here, a set of software tools is presented that implement the previously described Characteristic Attribute Organization System for both diagnostic identification and diagnostic-based classification. The software is publicly available from http://sarkarlab.mbl.edu/CAOS.

Introduction
DNA barcoding initiatives have, to date, relied heavily on distance based and computationally intensive tree-based methods for both diagnostic identification and later classification of new data. Alternatively, character-based methods can be used to identify classification rules based on an existing hierarchical organization, and then rapidly classify new data without requiring intensive phylogenetic approaches. The utility of a character based approach to DNA barcoding has been discussed in a theoretical context and in the context of DNA barcoding's relevance to classical taxonomy (Prendini 2005; DeSalle et al. 2005; DeSalle, 2006; Rubinoff, 2006a; 2006b; Williams and Ebach, 2006; Little and Stevenson, 2007). Character based approaches have also been shown to be feasible and effective in recent publications (Rach et al., 2008; Kelly et al., 2007). However, an operational approach to, and software for, character based DNA barcoding has been lacking.

There are two parts to developing a valid character based approach for DNA barcoding. First is the establishment of the "DNA barcodes." This aspect is often referred to as the "species identification" process, involving the identification of characters that are diagnostics for pre-described and pre-determined taxonomic units. This step is equivocally akin to discovering morphological diagnostics in classical anatomical taxonomy. After the character based DNA barcodes have been established (i.e., once DNA diagnostics have been identified), a DNA barcode "reader" needs to classify unknown query specimens based on the identified DNA barcode diagnostics. In the case of both distance and tree-based approaches, the DNA barcode is equivalent to the entire DNA sequence and the barcode reader considers the position of the query sequence in a phylogenetic tree. In contrast, a character-based approach considers ONLY the diagnostic sites and the barcode reader considers the number of diagnostic positions that are relevant for supporting a particular classification.

We have developed a downloadable software tool to implement a character-based approach to DNA barcoding. This software first identifies diagnostic DNA sequence changes in a data set and establishes those as "rules" for the second function of the program that can read DNA sequences from query specimens and identify them to their species. This program is based on the Characteristic Attributes Organization System (CAOS) algorithm described in (Sarkar et al., 2002a; 2002b). CAOS is an automated systematic method for discovering conserved character states from cladograms (i.e.,

trees) or groups of categorical information. CAOS defines attribute tests at each node in a phylogenetic tree, similar to decision tree algorithms. Character states, called "attribute tests" in decision trees, are termed "Characteristic Attributes" (CAs) in CAOS. Unlike decision tree algorithms, CAOS does not consider every possible attribute test. Instead, CAOS only considers diagnostically informative attributes. Figure 1 provides a graphical overview of diagnostic CA's that CAOS identifies. CAOS searches for two types of CA's: pure and private. Pure attributes (Pu) exist across all members of a single clade, and never in any other clade. Private attributes (Pr) are present across some of the members of a clade, but never in any other clade. CAOS first identifies all CA's that are comprised of single character states. These are called "Simple" CA's (sCA's). Simple Pure CA's are called sPu's; Simple Private CA's are called sPr's. CAOS then searches for multi-character states called "Compound" CA's (cCA's). cCA's are not composed of any other CA's (e.g., a collection of sCA's does not comprise a cCA). Compound Pure attributes and Compound Private attributes are designated as cPu and cPr, respectively. CA's are organized into an induction rule set that can be used to classify new sequence (a "taxon") onto the original tree wherefrom the CA's were identified. Similar to the decision tree methods for classification, CAOS classifies a new taxon by tracing a path of rules (i.e., the CA's) to the final classification state. Different combinations of CA's can be used to classify different taxa to the same node. This reflects the combination of the partitional method for classification of a taxon at each node within the context of the hierarchical classification of the dendrogram. Therefore, while the topology of the original tree is used as a guide for a decision tree, the decision at each node is determined using a partitional clustering-like algorithm. An overview of the entire CAOS process is shown in Figure 2.

The CAOS system consists of two programs: P-Gnome and P-Elf. P-Gnome is a diagnostic rules generator that searches through a given data matrix and establishes diagnostic rule sets for each of the pre-described entities in the data matrix. The P-Elf program can then classify a file of query sequences according to the rules generated by P-Gnome.

Download and system requirements

The program is downloadable from http://sarkarlab.mbl.edu/CAOS. The program is currently available as a Mac OS X installer (source code that can be compiled on other *NIX platforms is also available from http://phylogenomics.googlecode.com/svn/trunk/CAOS/).

P-Gnome and P-Elf are Perl scripts that interact with a CAOS engine that was written in C++. Perl is pre-installed on all Mac OS X systems since version 10.3; however, for older versions of OS X, Perl can be installed via the Apple Developer Tools (ADT), which comes with the OS X installations discs (or the ADT can be downloaded directly from Apple (http://developer.apple.com/tools). Perl is generally part of the default installation in most flavors of *NIX.

Input files and output interpretation

P-Gnome uses the NEXUS file format. P-Gnome uses a specific format of NEXUS files, which consists of a non-interleaved DNA data matrix, a translate block that converts the taxon names to integer values in the tree representation, and a Newick tree embedded in the file that is the result of collapsing nodes relative to the taxonomic groupings of interest (i.e., pre-described species boundaries). It is possible to manually create this file; however, the process is greatly facilitated through the use of phylogenetic examination packages like MacClade (Maddison and Maddison, 2005) or Mesquite (Maddison and Maddison, 2007). P-Gnome output is placed in several files of which the attributes and group files are most relevant for establishing diagnostics. The "groups" file interprets the input tree and places all terminals into nested groups that can be used or discarded as hypotheses of species grouping. The "attributes" file lists the diagnostics and level of confidence for the diagnostics for each of the groups established in the groups file. To better organize results from a CAOS analysis, the output from P-Gnome is designed and formatted such that they can be manipulated in spreadsheet applications, like Microsoft Excel. Based on the rule files generated by P-Gnome, the P-Elf script reads FASTA files and returns a file with the best identification for each input query either individually or in batch. This output is also simple to manipulate in spreadsheets. The Supplemental Manual explains the output in detail and sample matrices for testing CAOS operations are available online at http://sarkarlab.mbl.edu/CAOS.

Utility

The program is rapid; data sets with hundreds of sequences (e.g., Rach et al., 2008; Kelly et al., 2007) have been analyzed for diagnostics in less than ten minutes. In addition, there are two immediately important uses for the CAOS package. The first utility is to establish diagnostics for DNA barcoding and the second is useful as a DNA barcode reader. Scientists interested in DNA barcoding in taxonomy will find the first function useful to establish diagnostics from DNA sequences for species descriptions. Other areas of utility for CAOS are in ecology, biodiversity and conservation biology studies that can utilize the program as a DNA barcode reader to identify unknown query specimens relevant to these areas of biology. Scientists interested in species discovery (DeSalle, 2006; Rubinoff, 2006b) can use the program to test hypotheses of species existence through the application of the phylogenetic species concept as articulated by Davis and Nixon (1992). The CAOS program might also be useful for other approaches to species discovery that utilize non-phylogenetic species concepts by making available an organizing system for complex DNA sequence data.

Future Versions

The current versions of P-Gnome and P-Elf are designed primarily as command-line applications. We are in the process of developing a graphical (e.g., Web) interfaces to CAOS as well as a visualization module to view identified diagnostics within the context of DNA barcoding. The Web version, planned for late 2009, and all preceding future releases of the source code and compiled application binaries will be made available both from the CAOS website (http://sarkarlab.mbl.edu/CAOS) and the Google Code page (http://phylogenomics.googlecode.com/svn/trunk/CAOS/).

References

Davis, J. I. and K.C. Nixon. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. *Systematic Biology* 41: 421-435.

De Salle, R., M.G. Egan and M. Siddall. 2005. The unholy trinity: taxonomy, species delimitation, and DNA barcoding. *Phil. Trans. R. Soc. B*. 360 (1462) 1905 – 1916.

DeSalle, R. 2006. Species Discovery versus Species Identification in DNA Barcoding Efforts: Response to Rubinoff . *Conservation Biology*. 20 (5) 1545-1547.

Kelly, R.P., I.N. Sarkar, D.J. Eernisse and R. Desalle 2007. DNA barcoding using chitons (genus Mopalia) - *Molecular Ecology Notes*. 7, 177–183.

Little, D. P. and D. Wm. Stevenson. 2007. A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms. Cladistics 23 (1): 1–21.

Maddison, W. P. and D. R. Maddison. 2005. MacClade: Analysis of Phylogeny and Character Evolution. Version 3.0. Sinauer Associates, Sunderland, Massachusetts.

Maddison, W. P. and D.R. Maddison. 2007. Mesquite: a modular system for evolutionary analysis. Version 2.01 http://mesquiteproject.org.

Prendini, L. 2005. Comments on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology*. 83(3) 498-504(7).

Rach, J., R. DeSalle, I.N. Sarkar, B. Schierwater and H. Hadrys. 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Phil. Trans. R. Soc. B*.. 275(1632): 237-247.

Rubinoff, D. 2006a. Utility of Mitochondrial DNA Barcodes in species conservation. Conservation Biology 20:1026-1033.

Rubinoff, D. 2006b. DNA Barcoding Evolves into the Familiar . *Conservation Biology*. 20 (5) 1548-1549.

Sarkar I.N., P.J. Planet, T.E. Bael, S.E. Stanley, M.E. Siddall, R. DeSalle and D.H. Figurski. 2002a "Characteristic Attributes in Cancer Microarrays." *Journal of Biomedical Informatics* 35(2):111-122.

Sarkar I.N., J. Thornton, P.J. Planet, B. Schierwater and R. DeSalle. 2002b. "A systematic method for classification of novel homeoboxes." *Molecular Phylogenetics and Evolution* 24(3):388-399

Schindel, D. E. and S.E. Miller. 2005. DNA barcoding a useful tool for taxonomists. *Nature*. 435 17.

Will, K.W., B.D. Mishler and Q.D. Wheeler. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*. 54(5) 844 - 851.]

Williams, D.M and M.C. Ebach. 2006. Foundations of Systematics and Biogeography. Springer, New York.

FIGURE LEGENDS.

Figure 1: The Characteristic Attribute Organization System. Types of characteristic attributes (CAs) in a hypothetical set of DNA sequences are shown for a matrix of characters organized into two groups (each consisting of four taxa). Pure (Pu) CAs are character states that exist in all elements of a given group but not in any members of the alternate group; private (Pr) CAs are present in only some members of a group but are absent from the alternate group. Simple (s) CAs are attributes at single positions. Compound (c) CAs are combinations of states, that are not already considered CAs.

Figure 2: Overview of CAOS-based DNA barcoding workflow. The overall process for using the CAOS system for DNA barcoding is shown as a four-step process: (1) Acquire DNA molecular marker sequences (e.g., CO-I) and perform a phylogenetic analysis; (2) Collapse nodes as appropriate (i.e., at pre-defined species boundaries) and generate NEXUS file; (3) Determine diagnostic positions at each major taxonomic grouping using the CAOS P-Gnome program; (4) Classify new sequences into taxonomic groupings using the CAOS P-Elf program.

Figure 1.

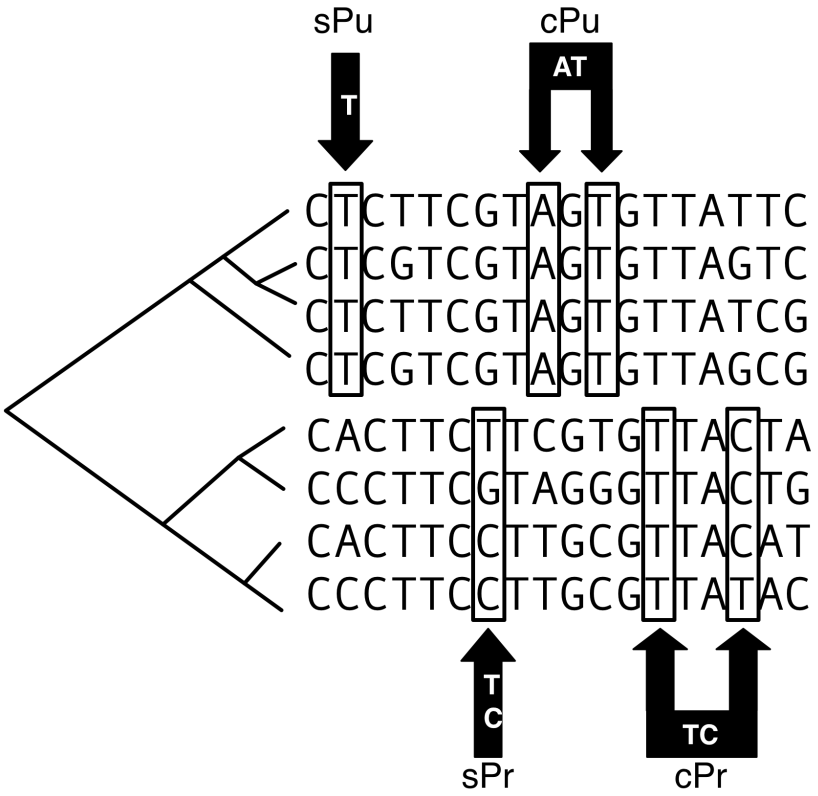
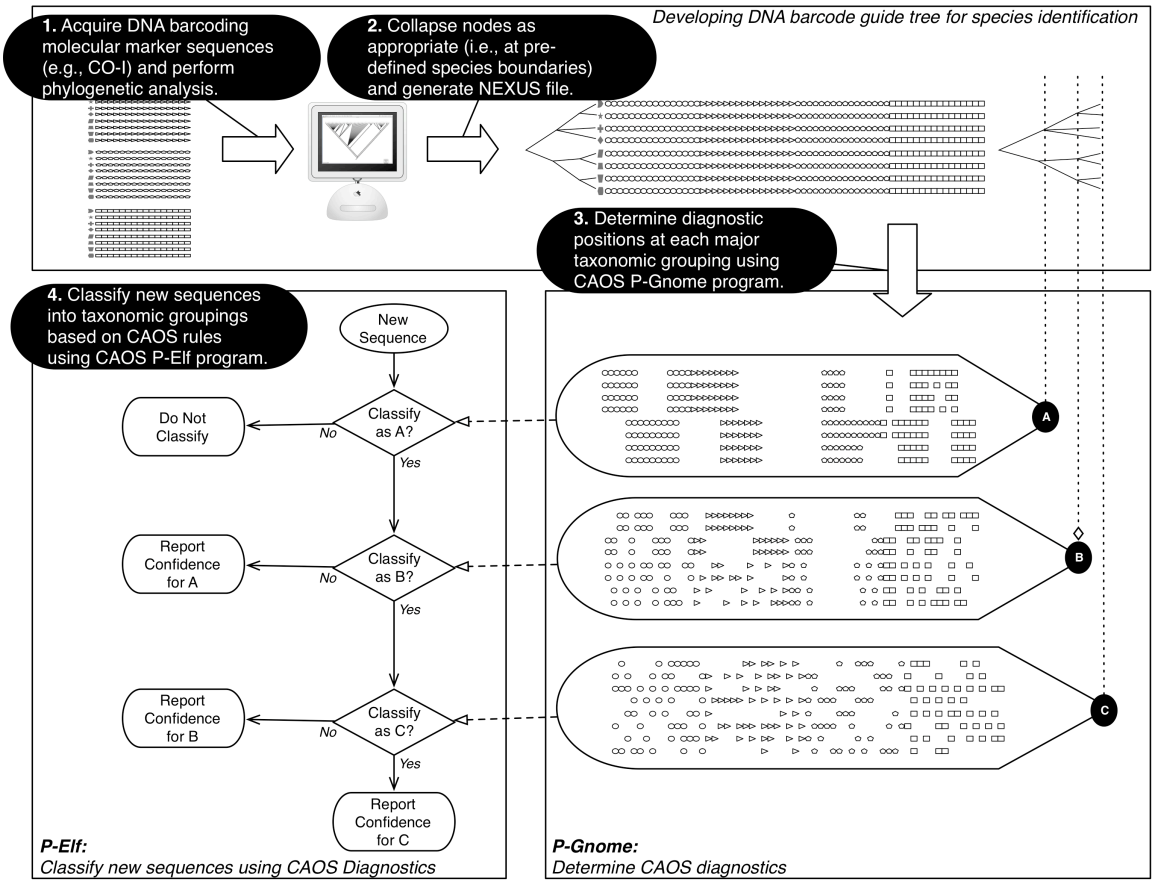

sPu
cPu
T
AT
CTCTTCGTAGTGTTATTC
CTCGTCGTAGTGTTAGTC
CTCTTCGTAGTGTTATCG
CTCGTCGTAGTGTTAGCG
CACTTCTTCGTGTTACTA
CCCTTCGTAGGGTTACTG
CACTTCCTTGCGTTACAT
CCCTTCCTTGCGTTATAC
T
C
TC
sPr
cPr

Figure 2.


Developing DNA barcode guide tree for species identification
1. Acquire DNA barcoding molecular marker sequences (e.g., CO-I) and perform phylogenetic analysis.
2. Collapse nodes as appropriate (i.e., at pre-defined species boundaries) and generate NEXUS file.
3. Determine diagnostic positions at each major taxonomic grouping using CAOS P-Gnome program.
4. Classify new sequences into taxonomic groupings based on CAOS rules using CAOS P-Elf program.
New Sequence
Classify as A?
No
Yes
Do Not Classify
Classify as B?
No
Yes
Report Confidence for A
Classify as C?
No
Yes
Report Confidence for B
Report Confidence for C
A
B
C
P-Elf:
Classify new sequences using CAOS Diagnostics
P-Gnome:
Determine CAOS diagnostics