1001 - A tool for binary representations of unordered multistate characters
(with examples from genomic data)

Mavrodiev E. V.¹

¹University of Florida, Florida Museum of Natural History, Gainesville, FL 32611 USA. evgeny@ufl.edu

Abstract. In modern molecular systematics, matrices of unordered multistate characters, such as DNA sequence alignments, are used for analysis with no further re-coding procedures nor any a priori determination of character polarity. Here we present 1001, a simple freely available Python-based tool that helps re-code matrices of non-additive characters as different types of binary matrices. Despite the historical basis, our analytical approach to DNA and protein data has never been properly investigated since the beginning of the molecular age. The polarized matrices produced by 1001 can be used as the proper inputs for Cladistic analysis, as well as used as inputs for future three-taxon permutations. The 1001 binary representations of molecular data (not necessary polarized) may also be used as inputs for different parametric software. This may help to reduce the complicated sets of assumptions that normally precede either Bayesian or Maximum Likelihood analyses.
Introduction

In an un-polarized, binary matrix the states 0 and 1 do not represent a hypothesis of character polarity. In an polarized binary matrix character-states 0 and 1 are considered to be plesimorphic (“primitive”) and apomorphic (“derived”) respectively apriori to analysis (see Kitching et al. 1998). Because in Cladistics groups must be define based only on the synapomorphies (e. g., Donoghue and Maddison 1986, Williams and Ebach 2008), it is critical to assume states’ polarity before analysis and group only on the states “1” of the polarized binary matrix (e. g., Platnick 1985, 2013, de Pinna, 1996, Williams and Ebach 2008, Waegele, 2004, 2005). Therefore a fundamental problem in molecular systematics today is that molecular matrices are not polarized (Waegele 2004, 2005) and are therefore analytically uninformative form Cladistics standpoint (Williams and Ebach 2008).

Historically, polarized binary matrices were proposed as an ideal data format for cladistics analysis following the argumentation schemes of Willi Hennig, who determined each character’s polarity before the construction of a cladogram (e. g., Swofford and Begle 1993, Kitching et al. 1998, Waegele 2004, 2005, Williams and Ebach 2008, Ebach et al., 2013, Wiley and Lieberman 2011). Hennigian logic may be clear even from the pure methodological standpoint: without a character hypothesis in place apriori to analysis, we are unable to test hypotheses aposteriori (Ebach, personal note).
Given this, we have developed 1001, a computer program that converts un-polarized molecular data matrices into the different types of binary matrices, either un-polarized or with established polarity.

Implementing 1001

1001 is implemented as Python-based script that translates conventional matrices of unordered multisite characters into polarized and non-polarized binary matrices written in Phylip or Comma Separated Values (CSV) file formats. 1001 can be used with any operating system that has a Python interpreter (e.g., Linux, Mac OS X, and Windows (http://www.python.org/). Script been written by Dr. Matthew A. Gitzendanner (University of Florida, Department of Biology and Florida Museum of Natural History, Gainesville, FL 32611 USA).

1001 accepts DNA and amino acid sequence alignments, as well non-molecular data, in “relaxed” PHYLIP format (e.g., Maddison and Maddson 2011). All gaps and ambiguities of the conventional multistate matrices must be recoded as "?" ("missing entities") before running of 1001. The DNA sequence data is the subject of our primary interest and the default option of 1001 designed for these kind of data. However, the script may handle different kinds of unordered multistate characters.

Figures 1 (A – C), 2 and 3 provide the summary and the explanation of the results and methods. Both methods implemented by 1001 a priori polarize conventional data by comparing with an assumed all-plesiomorphic outgroup, as it was proposed before for morphological data. This way of re-coding led to the non-direct methods of polarity

The first Method (Fig. 1B) is based on a standard bioinformatics technique frequently cited as the “Vos representation” of DNA sequences (reviewed in Bernaola-Galvan et al. 2002) or as “CODE-4 encoding” of DNA data (Demeler and Shou 1991: 1594, see also absence/presence coding of Carine and Scotland 1999, Scotland 2000a, b and Pleijel 1995 (“Method D”), reviewed in Kitching et al. 1998), but, additionally, with the re-coding of the resulted 1/0 matrix as the polarized binary matrix.

Eight output files resulted from each run of 1001, if the First Method selected:
- non-polarized binary matrix, with and without invariant characters removed (both phy and csv files);
- polarized binary matrix, with and without invariant characters removed (both phy and csv files);

The absence/presence coding may results the similar trees as obtained with the regularly coded characters, or may bias the original multistate data (de Laet 2005: 94-96) therefore an additional method of the binary representation of multistate data also implemented in 1001. Additional ways of binary coding are also possible, at least for the DNA characters (e. g., Bernaola-Galvan et al., 2002: 106, Table 1).

This second method (Fig. 1C), or as we prefer to call it the “Cladistic” Method, directly represents the conventional multistate alignment as a set of maximum
relationships (Williams and Ebach 2008) following the values of the pre-selected out-group taxon (Fig. 1). This method is designed for the polarized binary outputs only (available in both phy and csv formats).

Both proposed methods increase the number of parsimony-informative characters.

Both polarized and non-polarized binary 1001 outputs can be used with popular phylogenetic software, with many different statistical packages as well as an input for three-taxon permutations (Nelson and Platnick 1991) using TAXODIUM 1.2 (polarized binary data only)(Mavrodiev and Madorsky 2012) for the future completion of the Cladistics analysis.

1001 is available for free from the Web (https://github.com/magitz/1001)

**Breaking with the tradition of using unpolarized matrices in molecular systematics**

Many popular phylogenetic applications are able to polarize characters before analysis (e.g., command “AncState” of PAUP* (Swofford 2002) and “ancstates” of TNT (Goloboff et al. 2008), see also the option “ancestors” included in some programs of PHYLIP package (Felsenstein 1989). However our analytical approach to DNA and protein data has never been properly investigated since the beginning of the molecular age (Waegele 2004, 2005).

explicitly did not recommended to polarize characters before analysis, and was initially strongly criticized for this position (Donoghue and Maddison 1986).

However, as was later clearly summarized by Swofford and Begle (1993: 3, 27) in general agreement with Meacham (1984, 1986), it is better to infer the topology of the tree and the character polarities simultaneously, rather than going through the two-stage process of assigning polarities first and then estimating the tree (see also Maddison et al., 1984, among others).

Maddison et al. (1984) and Swofford and Begle (1993) also noted that a priori polarization of the characters is reasonable only when the polarity determination is unambiguous (i.e., there is no heterogeneity in the outgroup for characters that are variable within the ingroup); when the outgroup is heterogeneous, and the most parsimonious assignment of an ancestral condition for the ingroup depends upon how the outgroup taxa are related to each other (Swofford and Begle 1993: 3, 27, see also Maddison et al., 1984, Nixon and Carpenter, 1993, Maddison and Madison, 2011, Kitching et al. 1998, Lyons-Weiler et al. 1998 among others).

In other words, if the number of taxa within the out-group is in some way reduced to one (see Arnold 1989 among others including the TNT program (Goloboff et al. 2008) that offer a single out-group taxon as a default option) (or also in the case of homogeneous outgroup), for a character with two or more states, the state occurring in the outgroup can be indeed assumed to be the plesiomorphic state (Platnick and Gertsch 1976: 2, see also Watrous and Wheeler, 1981, Bryant 2001, de Pinna 1994, Kitching et al. 1998, Donoghue and Maddison 1986, and Nixon and Carpenter 1993 for the reviews).
We believe that numerous analytical possibilities are still missed from this simple cladistics perspective and therefore the 1001 may help to investigate the field better.

*If the characters of conventional multistate matrix are polarized, then the data are represented in the form of relations*, either “maximum” or “minimum” (Williams and Ebach 2008). One of the goals of 1001 is the explication of sets of “maximum relationships” as separate entities (as polarized binary matrices) for future analyses. The listed popular software (see above) is unable to perform such explications, even if in principle these programs can polarize matrices before analyses.

Each relation is a not equal to the conventional character anymore, but represents the hypothesis of the relationships between taxa (e. g., Platnick *et al.* 1996, Williams and Ebach 2008). Therefore the polarized binary matrix is semantically different from either raw multistate alignment or from the non-polarized binary representation of this alignment. One, for example, may note that the polarized binary matrix represents a kind of *structure*, rather than the collection of raw characters.

Another may tell us that the notion that systematic data constitutes a normal character by taxon matrix is not an intrinsically cladistic notion (Platnick 1993: 271, see also Williams and Ebach 2006, 2008 and Ebach *et al.* 2013) and, therefore, another type of data may require for the Cladistic analysis, especially if the last one is viewed as an extension of comparative approach (e. g., Nelson and Platnick 1981, Williams and Ebach 2008, Rieppel *et al.*, 2013, see also Nelson, 1970). The sets of maximum relationships explicated by 1001 may be considered as putative candidates for the proper inputs for Cladistic analysis.
1001 and three-item analysis

It was argued multiple times, that the three-taxon matrix constitutes the actual systematic data (e. g., Platnick 1993: 271, see also Nelson and Platnick 1991, Williams and Siebert 2000 and Williams and Ebach 2008). However the three-taxon representation of unordered multistate data may be an issue (Williams and Siebert 2000).

Scotland and others (Carine and Scotland 1999, Scotland 2000a, b) already performed the three-taxon-permutations of non-additive binary data (eventually re-coded multistate data). They also discussed their methodology in the context of Patterson’s idea of “pair homology” (Carine and Scotland 1999, Scotland 2000a, b). The polarized binary outputs of 1001 may also be used as inputs for future three-taxon representations using TAXODIUM 1. 2 (Mavrodiev and Madorsky 2012)(Fig. 2). As well as the Williams - Siebert three-taxon representation of unordered multistate data (Williams and Siebert 2000, Mavrodiev and Madorsky 2012, Mavrodiev et al. 2014), this option may help to prevent the artificial groupings under the conditions of 3TA, mentioned by Kluge and Farris (1999) and Farris et al. (2001) in their comments of the results of Scotland and others (Carine and Scotland 1999, Scotland 2000a, b).

1001 and parametric methodology

Below we argued that it is critical to break with the tradition of using unpolarized matrices in molecular systematics. However even the non-polarized 1001-binary representations of molecular data may essentially extend the horizons of conventional
phylogenetic analyses. For example these representations may also be used as inputs for parametric software and analyzed under the conditions of the simplest Mk model (Lewis, 2001) and it elementary binary derivates (Stamatakis 2014)(Fig. 3). This may help to simplify the complicated sets of assumptions (Jefferys and Berger 1992, Berger and Jefferys 1992) that normally precede either Bayesian or Maximum Likelihood approaches increasing the “robustness” of the analyses (Berger and Jefferys 1992). More investigation is necessary here, however.

Conclusion

A fundamental problem in molecular systematics today is that molecular matrices are not polarized. Historically, specifically polarized binary matrices were been proposed as an ideal data format for Cladistic analysis following the argumentation schemes of Willi Hennig, but despite of historical background, this analytical approach to the molecular data, never been properly investigated. Given this, we have developed 1001, a simple computer program that converts un-polarized molecular data matrices into the different types of binary matrices, either un-polarized or with established polarity. Both methods implemented by 1001 a priori polarize conventional data by comparing with an assumed all-plesiomorphic outgroup, as it was proposed before for morphological data. The polarized matrices explicated by 1001 may be considered as candidates for the proper inputs for Cladistic analysis, as well as used as inputs for future three-taxon permutations. The 1001 binary representations of molecular data (not necessary polarized) may also be used as inputs for different parametric software. This may help to reduce the complicated
sets of assumptions that normally precede either Bayesian or Maximum Likelihood analyses.
Acknowledgments

I thank Prof. Gareth Nelson for brief helpful discussion regarding the Cladistics representation of the DNA characters (Method 2), as implemented in “1001”. I also thank Dr. Matthew Gitzendanner for elegant Phyton-based implementation of “1001”. Finally, I would like to acknowledge Drs. David Williams and Malte Ebach for their helpful comments and linguistic remarks that help to sharp arguments. No agreement imply from the behalf of any person acknowledged in this section.
References


Figure Legends

Fig. 1. A. Results of Maximum Parsimony Analyses (MP) of conventional plastid genomic DNA matrix from Ma et al. 2014a, b. Final topologies rooted relatively *Dendrocalamus latiflorus* Munro (Ma et al. 2014). Median Consensus Tree based on Robinson-Foulds (RF) distance (with the best score found = 8837) of 184 shortest trees of length = 5019 (CI = 0.89, RI = 0.91). Number of taxa = 157. All constant characters from the original alignment are excluded from the analysis. Number of variable characters = 4304, number of parsimony-informative characters = 2003. * nodes received Maximum Parsimony Jackknife (JK) support >50% after 20 000 fast JK replicates; ! nodes recovered Maximum Parsimony Bootstrap support in the original analysis of Ma et al. 2014 (200 full heuristic replicates). B. Results of Maximum Parsimony Analyses (MP) of binary representation of conventional DNA matrix from A., re-coded following proposed Method 1. Initial binary date were also polarized before analysis relatively *D. latiflorus*, assumed as an outgroup based on the results of Ma et al. (2014a). Majority-Rule Consensus of 191 shortest trees of length = 10014 (CI = 0.88, RI = 0.89). Number of taxa = 157. Number of binary characters = 8783, number of parsimony-informative characters = 4088. * nodes received Maximum Parsimony Jackknife (JK) support >50% after 20 000 fast JK replicates. C. Results of Maximum Parsimony Analyses (MP) of binary representation of conventional DNA matrix from A., re-coded following proposed Method 2. Data polarized before analysis relatively *Dendrocalamus latiflorus*, assumed as an out-group based on the results of Ma et al. 2014. Majority-Rule Consensus of 139 shortest trees of length = 4993 (CI = 0.89, RI = 0.91). Number of taxa = 157. Number of
binary characters = 4993, number of parsimony-informative characters = 2027. * nodes received Maximum Parsimony Jackknife (JK) support >50% after 20 000 fast JK replicates.

All MP analyses were conducted using program PAUPrat (Nixon 1999; Sikes and Lewis 2001; Swofford 2002) as implemented in CIPRES (Miller et al. 2010) following 200 ratchet replicates with no more than 1 tree of length greater than or equal to 1 saved in each replicate, and the TBR branch swapping/MulTrees option in effect; -pct = 20%, all characters weighted uniformly, gaps were treated as “missing”. Maximum Parsimony jackknifing (Farris et al. 1996) conducted using program PAUP* (Swofford 2002). Robinson-Foulds consensus (reviewed in Kitching et al. 1998 and Bansal et al. 2010) calculated using program RFS v. 2.0 (Bansal et al. 2010). Majority-Rule consensus calculated in PAUP* (Swofford 2002).

All gaps and ambiguities of the conventional DNA matrix (A.) recoded as missing data (“?”) before binary permutations.

Roman numbers corresponds to the “major lineages” of bamboos, as specified by Ma et al. 2014a.

Fig. 2. The results of two preliminary three-taxon analyses (3TAs) of Clades 1 and 2 of the general topology, described on Fig. 1. In both cases, the DNA alignments derived from described above (Fig. 1, all original data came from Ma et al. 2014a, b) simply by proper sampling. These alignments been polarized following Method 2 and after that established as a tree-taxon matices using TAXODIUM 1.2 (Mavrodiev, Madorsky, 2015). Indocalamus wilsonii (Rendle) C.S.Chao & C.D.Chu (Clade 1) and Bergbambos
tessellata (Nees) Stapleton (Clade 2) assumed as an outgroup taxa before Method 2 applied to the DNA characters. A. The results of the first 3TA (Clade 1). Majority-Rule Consensus of 193 shortest trees of length = 527046 (CI = 0.92, RI = 0.91). The number of taxa in 487168 character-3TA matrix is 72. All 487168 3TSs are parsimony-informative and weighted uniformly. B. The results of the second 3TA (Clade 2). Majority-Rule Consensus of 201 shortest trees of length = 187857 (CI = 0.86, RI = 0.83). The number of taxa in 161027 character-3TA matrix is 80. All 1610278 3TSs are parsimony-informative and weighted uniformly. Roman Numbers corresponds to the “major lineages” of bamboos, as specified by Ma et al. 2014a. See also the Legend of the Fig. 1 for the details of the MP analyses.

**Fig. 3.** A. Most probable topology recovered from a Maximum Likelihood (ML) analysis (RAxML)(Stamatakis 2014) of conventional *Campanula* s.l. & outgroups DNA plastid + nuclear (PPR loci) combined matrix from Crowl et al. (2014). The names of the all taxa are taken from the Supplemental data of Crowl et al. (2014). ML bootstrap (BS) values for nodes receiving .50% supports are indicated above and below the branches (1000 rapid replicates). GTR + G model was assumed to be the best choice for the molecular dataset. Final ML Optimisation Likelihood: -57876.127159. B. Most probable topology recovered from Maximum Likelihood analysis (RAxML)(Stamatakis 2014) of conventional *Campanula* s.l. & out-groups DNA plastid + nuclear (PPR loci) combined matrix from Crowl et al. (2014), but established as a non-polarized binary matrix following Method 1. Binary matrix analyzed under the assumptions of BINGAMMA model (Stamatakis 2014, see also Lewis 2001) with the putative ascertainment bias
(Lewis 2001) left uncorrected. ML bootstrap values for nodes receiving .50% supports are indicated above and below the branches (1000 rapid ML BS replicates). Final ML Optimisation Likelihood: -85930.633845.
A. DNA characters

B. Binary data: Method 1
(polarized binary example)

C. Binary data: Method 2

Fig. 1
Fig. 1 (continous)
Fig. 2
Fig. 3