Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

New measures of topological stability in phylogenetic trees – Taking taxon composition into account

Dirk Krüger [1,2*] and Andrea Gargas [1]

[1]University of Wisconsin, Botany Department, 430 Lincoln Drive, Madison, WI 53706, USA; [2]University of Wisconsin, Department of Soil Science, 1525 Observatory Drive, Madison, WI 53706, USA;

Dirk Krüger* - Email: dirkkrueger@wisc.edu; * Corresponding author

received December 27, 2006; accepted December 28, 2006; published online December 28, 2006

Abstract:

In phylogenetic trees the addition and removal of taxa has large effects on tree topology, hence measures of branch support and tree stability should account for taxonomic composition. Currently no comprehensive system of composition-dependent parameters exists in any cladistic or phenetic strategy. We introduce several values and indices based on a modification of the original jackknife resampling. Their advantage is a complete evaluation and optimization of taxon composition in phylogenetic data. While related to the Jackknife Monophyly Index (JMI), our system of support measures expands beyond parsimony analyses, and includes indices estimating support for the entire phylogenetic tree based on individual branch supports.

Keywords: phylogeny; tree topology; taxon composition; jackknife

Background:

Phylogenetics relies on statistical measures of tree robustness, branching topology, and ultimately of the data consistency. Stability of a branch location can be gauged under a stepwise relaxation of the optimality criterion selected. For example, in parsimony, the tree's sum of changes is the optimized criterion. Decay values on branches show how many changes are needed to cause a branch to disappear. **[1]** Confidence in a branch more commonly is estimated in a frequentist manner, entailing the occurrence of a particular branch in permuted or randomized data. This requires lengthy optimization of an optimality criterion until the solutions are summarized in a consensus tree. Tree stability is here estimated by altering the number and weight placed on homologous characters. At other times, frequentist measures are derived among equally supported solutions in a consensus, even in absence of data manipulations. Bayesian posterior probabilities are gathered from subsets of solutions after the likelihood criterion is optimized in the burn-in of Markov chain Monte Carlo. **[2]**

Common frequentist tree topology estimates are derived from random resampling with replacement of data (bootstrap) **[3]** or without data replacement (jackknife). **[4]** Each new data set is undergoing analysis. The frequency for a given branch to occur among those analyses is recorded. A consensus of the solutions on all pseudoreplicated data sets is generated, and includes the most frequent among competing branch locations. Usually a 50% minimum branch support is required for inclusion in the consensus tree. Often, a 95% interval is seen as significant support. Bayesian posterior probabilities and the Bayesian consensus tree are calculated in the same way, but based on a subset of retained optimized solutions. Posterior probabilities and non-parametric bootstrap values have different meanings and sensitivities to error. **[5]** The meaning of the bootstrap value in regards to phylogenetic accuracy is still debated. **[6, 7]**

The composition of taxa in a data set influences the tree topology **[4, 8]** as much as the composition of positionally homologous data characters. "Rogue taxa" strongly influence tree topology far beyond their own position in a tree. Long-branch attraction is the phenomenon where the most diverse taxa appear in a single clade. **[9]** Sometimes phylogeneticists take out controversial taxa implicitly, or explicitly, to estimate influence on the phylogeny. **[10]** Given that taxon sampling is important, measures of branch stability should include this aspect as well as those relying on character resampling. This has been taken up by Siddall **[11]** who invented a parameter called Jackknife Monophyly Index (JMI). Siddall's estimator was introduced for parsimony analyses. We take his approach a step further, which brings about a family of new values of support for tree branches, and overall tree stability. Our indices can be of further use in becoming optimality criteria, if one wants to find the optimal taxon composition for a given data set.

Methodology:

Three sample data sets

To compile three simple data sets, we used ITS region rDNA sequences from 6 puffball mushrooms. **[10]** Querying the public databases, we recorded pairwise BLAST scores from hitting each of the sequences in our data set. The scores where then converted to distance values by dividing BLAST scores by the maximum BLAST score in the matrix, and subtracting that value from 1 (distance = 1- BLAST score / maximum BLAST score).

Additionally, uncorrected p-distances (ClustalX) **[12]** where derived from sequence alignment using Divide-and-Conquer in QAlign. **[13]** The two resulting distance matrices were analyzed using UPGMA in the neighbor.exe module of PHYLIP. **[14]**

In a third analytical strategy, the alignment was submitted to Maximum-Parsimony (MP) non-parametric jackknifing (= jackknifing across genes, JAG) in DAMBE. **[15]**

Jackknifing across taxa (JAT)
We then manually removed from the distance matrices one taxon at a time, resulting in 6 new data matrices per strategy. The UPGMA analysis was then repeated on those. For the sequence alignment, we also removed one taxon at a time, and used the 6 new alignments with one of the original taxa missing each for MP analysis.

The JAT support value is defined by us as the frequency an internal branch (one that is not leading to a single taxon / tip) appears, as a percentage out of the total number of times it could appear. The latter takes into account that with a taxon missing due to resampling, that clade is no longer the same clade as with the taxon in. The JAT values are then assigned on the corresponding branches of the total tree solution of the data set, the one where all taxa are included. Zero-length terminal branches cause the underlying branch to not be considered as internal branch, as that branch in effect collapses. The iJAT index is defined as the sum of all JAT values shown on the total tree, divided by the number of JAT values for all potential clades.

JATxJAG
JATxJAG branch support values are shown based on the parsimony analysis. They are the product of the JAT support and the JAG support of a given internal branch, with iJATxJAG the overall support index of the total tree.

iJACK
This index is defined as the product of all traditional jackknife (resampling of characters) values in a tree. Calculated for each taxon subsampling separately, it can be shown as the average of all iJACK indices among the trees with different taxa removed.

Discussion:
Jackknifing across taxa can be modified to a random subsampling of more than one taxon at a time; such would be a higher-order jackknifing, e.g. with a random 50% of taxa deleted per pseudoreplicate. Similarly, all the conventions introduced here can be used in bootstrapping across taxa (BAT), with iBAT, iBOOT, BATxBAG, and iBATxBAT values. Theoretically, software could be implemented to optimize the iJAT, iJATxJAG, iJACK, and appropriate bootstrapping values by altering taxon composition among a large number of taxa available for the total data set.

The introduced measures of tree support are related to Siddall's Jackknife Monophyly Index. **[14]** Siddall conceived this stability indicator for parsimony analyses. A particular clade was seen as congruent even when one taxon was deleted, but all other taxa remained part of that clade in a pseudoreplicate analysis. The JAT values are more conservative. If taxon C is removed in a pseudoreplicate, one could not say if it would be inserted in one bipartition containing C, D, E, and F, or containing A, B, and C in a tree using all taxa. The JMI would allow counting any pseudoreplicate with clade D+E+F and any A+B as one also containing C. In other words, C could be counted as included with any branch.

In large data sets, neither increasing taxon number nor gene number **[8]** may ultimately allow inference of the true tree, so that researchers will still have to rely on resampling statistics to gauge data consistency. We have here introduced some new ways to do so.

The different support values corroborate each other, as shown for the JMI by Siddall. **[11]** Most difficult is the position of *Morganella subincarnata* (taxon F), in MP sitting on a short branch adjacent to D+E that also received low jackknife support, JAT, JAG, JAGxJAT. In the distance / UPGMA analyses, the bipartition D+E+F vs. A+B+C is not supported at all.

Conclusion:
The proposed modifications of the jackknife provide a simple way to measure the influence of taxon composition on branch stability when conducting phylogenetic analysis. Resampling characters as performed in the popular bootstrap method is in no way the only desirable option in estimating consistency of data. Extensions of branch supports into overall tree stability indices such as the iJAT may provide the basis for optimizable parameters for the detection of rogue taxa in complex data sets. These would point out taxa needing corroboration, explanation of unique evolutionary rates, or breaking of long-branch attraction.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

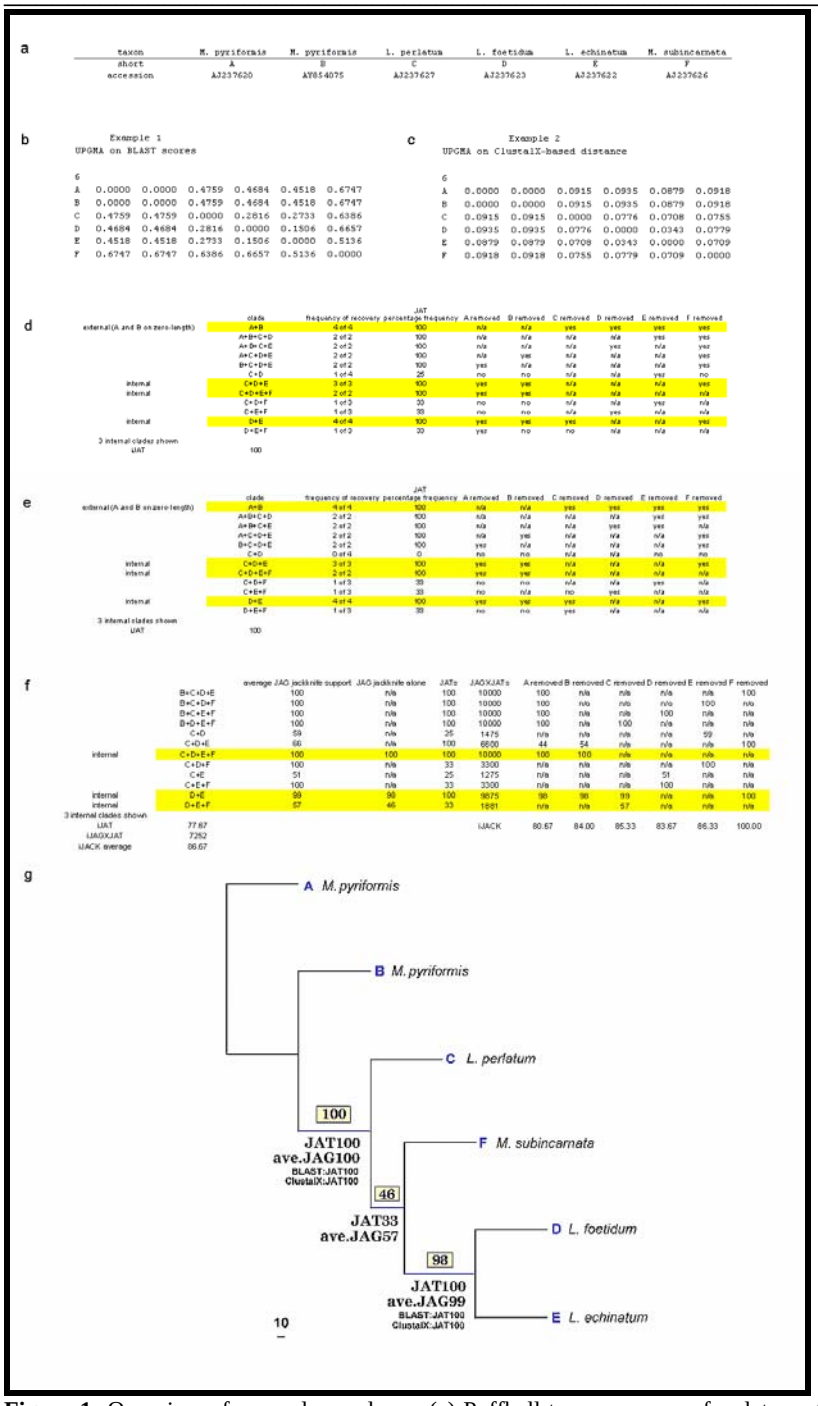

Figure 1: Overview of exemplar analyses. (a) Puffball taxa, acronyms for data matrices, and GenBank accession numbers (b) Distance matrix based on BLAST scores. (c) Distance matrix from ClustalX. (d) Calculation of JAT scores (= frequency of occurrence of internal clades) based on Figure 1b. (e) Calculation of JAT scores based on Figure 1C. (f) MP-related scores. (g) Composite phylogenetic tree based on 100 MP jackknife (JAG) pseudoreplicates, with new scores indicated below branches

Acknowledgement:
Antonis Rokas (Broad Institute, Cambridge MA) is thanked for early discussion of the idea, and Mirna Santana (University of Wisconsin, Madison WI) for helpful comments.

References:

[01] K. Bremer, *Cladistics*, 10:295 (1994)
[02] B. Rannala & Z. Yang, *J. Mol. Evol.*, 43:304 (1996) [PMID: 8703097]
[03] J. Felsenstein, *Evolution*, 39:783 (1985)
[04] S. Lanyon, *Syst. Zool.*, 34:397 (1985)
[05] M. P. Cummings, *et al.*, *Syst. Biol.*, 52:477 (2003) [PMID: 12857639]
[06] J. M. Carpenter, *Cladistics*, 8:147 (1992)
[07] M. J. Sanderson, *Syst. Biol.*, 44:299 (1995)
[08] A. Rokas & S. B. Carroll, *Mol. Biol. Evol.*, 22:1337 (2005) [PMID: 15746014]
[09] J. Felsenstein, *Syst. Zool.*, 27:401 (1978)
[10] D. Krüger, *et al.*, *Mycologia*, 93:947 (2001)
[11] M. E. Siddall, *Cladistics*, 11:33 (1995)
[12] J. D. Thompson, *et al.*, *Nucleic Acids Res*., 25:4876 (1997) [PMID: 9396791]
[13] M. Sammeth, *et al.*, *Bioinformatics*, 19:1592 (2003) [PMID: 12912847]
[14] J. Felsenstein, *PHYLIP* (Phylogeny Inference Package) version 3.5c, Univ. of Washington, Seattle (1993)
[15] X. Xia & Z. Xie, *J. Heredity*, 92:371 (2001) [PMID: 11535656]

Edited by P. Kangueane
Citation: Krüger & Gargas, Bioinformation 1(8): 327-330 (2006)