BAYES FACTOR FOR DATA COMBINABILITY

Application to the phylogeny of Sphaeropleales (Chlorophyceae, Chlorophyta)

Paul O. Lewis, Karolina Fučíková, and Louise A. Lewis

(Department of Ecology & Evolutionary Biology, University of Connecticut)

INTRODUCTION

Table 1. Data.

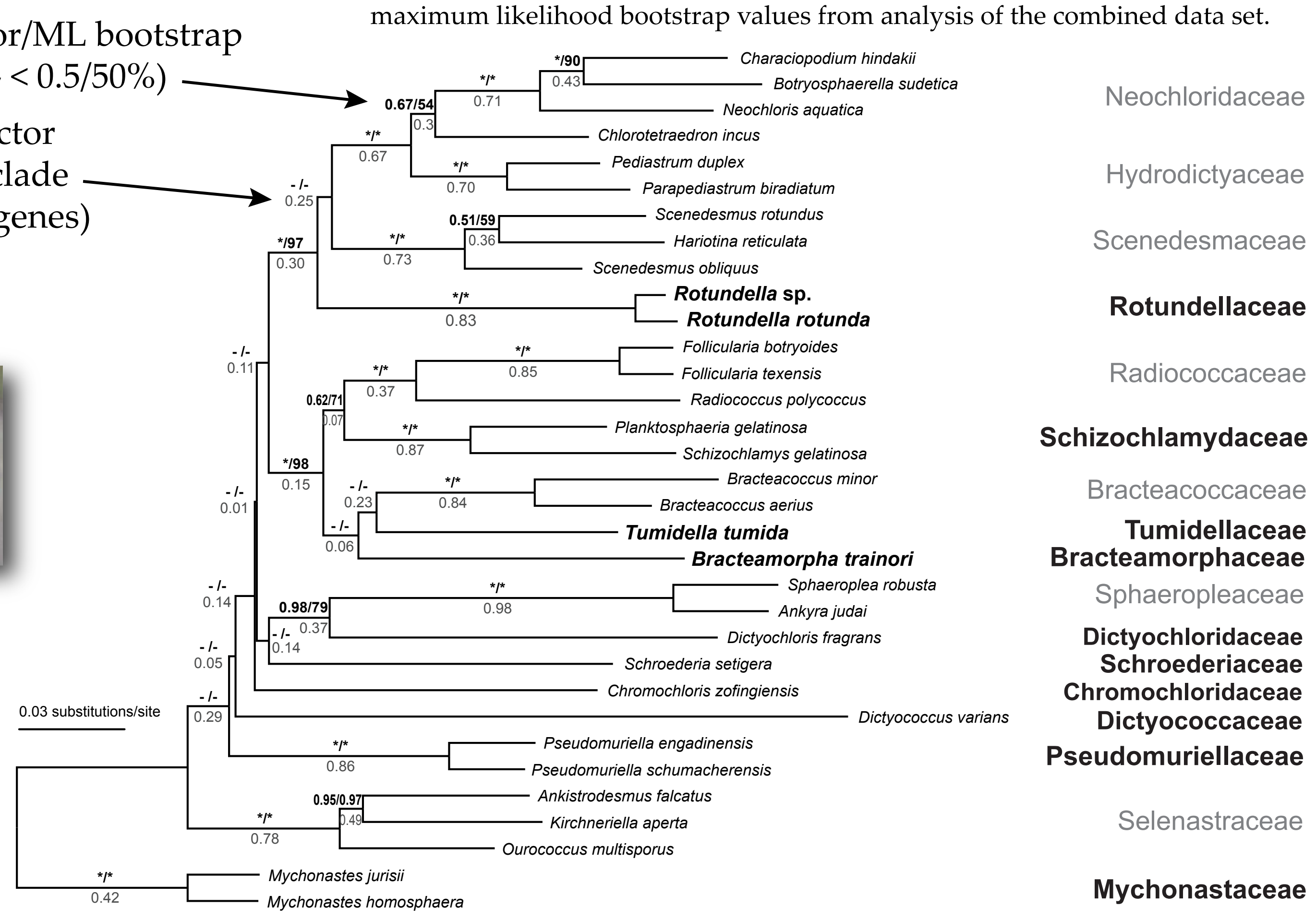
Ribosomal:
- 18S
- 28S
- 5.8S

Plastid:
- rbcL
- tufA
- psaB
- psbC

Bayesian posterior/ML bootstrap (*1.0/100%, - < 0.5/50%)

Concordance factor (0.25 means this clade shared by 25% of genes)

Figure 1. Primary Concordance Tree showing Bayesian posterior probabilities and maximum likelihood bootstrap values from analysis of the combined data set.



Tumidella tumida

Bracteamorpha trainori

Rotundella rotunda

The Problem

Fučíková et al. conducted separate Bayesian MCMC analyses of 7 genes (Table 1) in Sphaeropleales. Every distinct tree sampled was unique to one gene, causing Bayesian Concordance Analysis (BCA; Ané et al. 2007) to conclude maximum discordance. The BCA concordance tree (Figure 1) is (in this case) identical to a majority rule consensus of the pooled trees from all 7 single-gene analyses. Many clades were common to >70% of genes, suggesting there is more concordance than BCA indicates. Using results from only separate gene analyses precludes learning whether a single tree topology can adequately fit data from multiple genes.

Our Solution

We used Bayes Factors (BF) to assess combinability, taking advantage of a new method for estimating marginal likelihoods when tree topology is variable (Holder et al.). Our BF approach compares the probability of the data (marginal likelihood) when data subsets are combined (and thus forced to have the same tree topology) to the likelihood when subsets are separate (each having potentially a unique tree topology). While BCA suggests that no two genes can be combined, BF results favor combining most genes, excluding only 28S.

METHODS

The **Bayes Factor (BF)** is a ratio of marginal likelihoods computed under competing models.

BF > 1 means that model on top fits data better on average

BF < 1 means model on bottom fits data better on average.

Usually, a log scale is used for BF. Thus, log(BF) > 0 favors model on top.

Marginal likelihood of combined data

$$\mathrm{BF} = \frac{p(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_k)}{p(\mathbf{y}_1)\; p(\mathbf{y}_2)\; \cdots\; p(\mathbf{y}_k)}$$

Marginal likelihood of separate data

The **likelihood** is the probability of the observed data given a model.

Models include unknowns (parameters) such as tree topology, nucleotide frequencies, relative rates of substitution, branch lengths, etc.

The **marginal likelihood** is a weighted average of the likelihood over all combinations of unknown parameters, where weights are provided by the joint prior distribution.

GTR+I+G Substitution Model

6 exchangeabilities (5 free)

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & — & a\,\pi_C & b\,\pi_G & c\,\pi_T \\ C & a\,\pi_A & — & d\,\pi_G & e\,\pi_T \\ G & b\,\pi_A & d\,\pi_C & — & f\,\pi_T \\ T & c\,\pi_A & e\,\pi_C & f\,\pi_G & — \end{array}$$

4 relative frequencies (3 free)

1 proportion of invariable sites

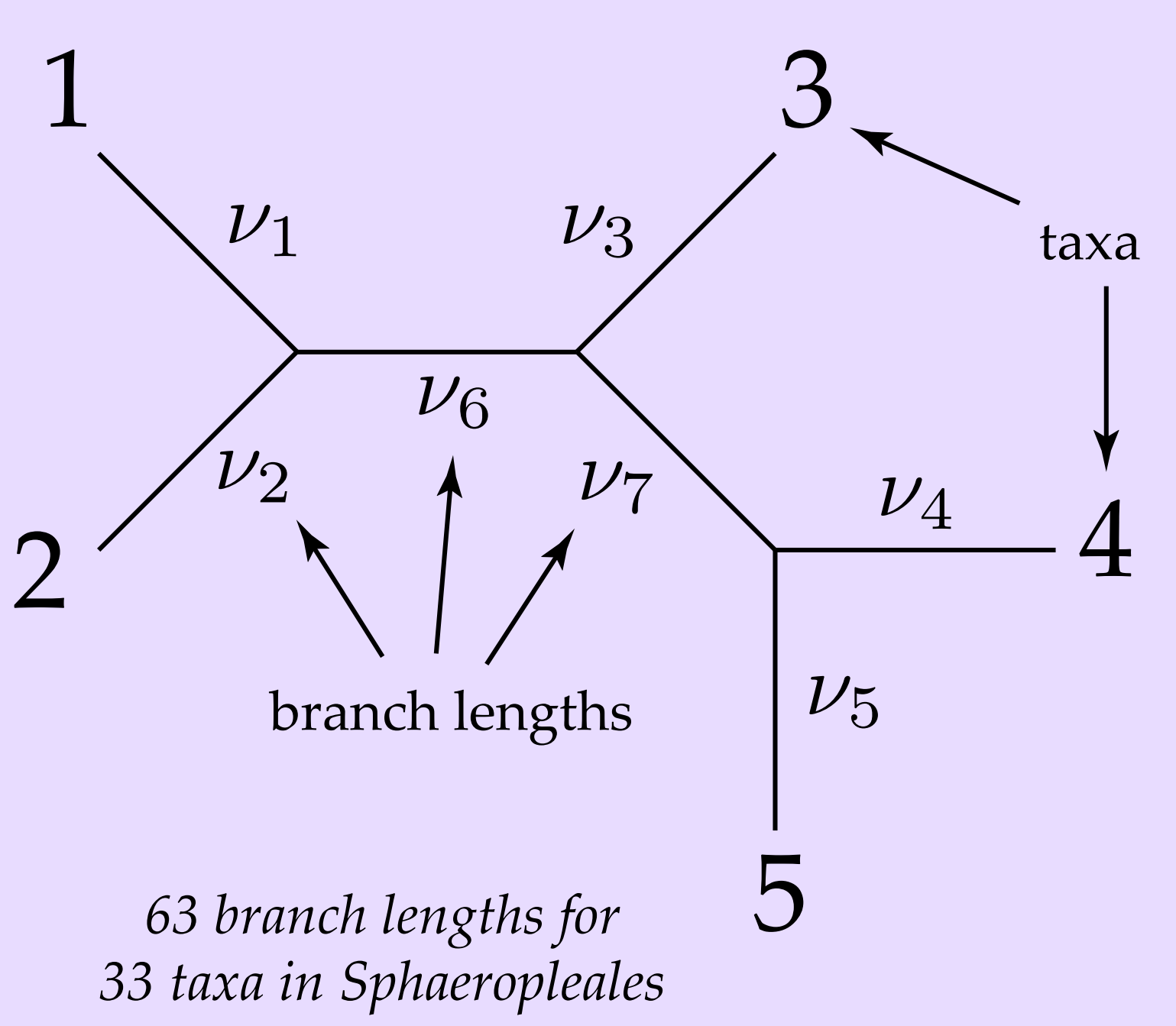
1 discrete gamma shape (among-site rate heterogeneity)

10 subst. model parameters for each partition subset

Partitioning

- Protein-coding data always partitioned by codon position
- Ribosomal data always unpartitioned
- Separate data sets never constrained to share tree topology or branch lengths

Tree Model



63 branch lengths for 33 taxa in Sphaeropleales study.

RESULTS

a. Ribosomal genes compatible: log(BF) = 67.5

Data subsets	Marginal likelihood	Number of free parameters
[28S 18S 5.8S]	**-17757.0**	(73)
[28S] [18S] [5.8S]	-17824.5	(219)

b. Plastid genes compatible: log(BF) = 132.9

Data subsets	Marginal likelihood	Number of free parameters
[rbcL tufA psaB psbC]	**-49096.2**	(93)
[rbcL] [tufA] [psaB] [psbC]	-49229.1	(279)

c. All genes compatible: log(BF) = 67.5
(but note 5.7:1 ratio of parameters)

Data subsets	Marginal likelihood	Number of free parameters
[28S 18S 5.8S rbcL tufA psaB psbC]	**-66971.9**	(103)
[28S] [18S] [5.8S] [rbcL] [tufA] [psaB] [psbC]	-67053.6	(591)

d. Plastid and ribosomal incompatible: log(BF) = -118.7

Data subsets	Marginal likelihood	Number of free parameters
[28S 18S 5.8S rbcL tufA psaB psbC]	-66971.9	(103)
[28S 18S 5.8S] [rbcL tufA psaB psbC]	**-66853.2**	(166)

e. Combinability when 1 gene left out

28S	18S	5.8S	rbcL	tufA	psaB	psbC	
●	●	●	●	●	●	○	✗
●	●	●	●	●	○	●	✗
●	●	●	●	○	●	●	✗
●	●	●	○	●	●	●	✗
●	●	○	●	●	●	●	✗
●	○	●	●	●	●	●	✗
○	●	●	●	●	●	●	✓

BF favors combining all 3 ribosomal genes (**a**) and all 4 plastid genes (**b**). BF also favors combining all 7 genes (**c**), so any incompatibility between plastid and ribosomal genes is apparently not strong enough to overcome the dimension penalty resulting from the fact that the separate model has 5.7 times more parameters than the combined model.

Incompatibility reveals itself when comparing plastid to ribosomal (**d**), where the ratio of parameters is only 1.6. Further experiments (**e**) reveal that, of the ribosomal genes, 28S appears to be most incompatible with the plastid genes. Panel (**e**) shows that leaving out 28S from the rDNA group allows the rDNA subset to be combined with the plastid subset.

DISCUSSION

Phylogenetic trees for different genes (gene trees) can differ from each other and from the species tree due to factors such as incomplete lineage sorting and horizontal gene transfer. Even if the true gene trees are identical, estimated trees can differ due to model misspecification leading to (for example) long branch attraction in some gene trees and not others.

If different genes evolved along the same tree, the data for these genes can be safely combined to increase the information available for estimating the phylogeny. Bayesian Concordance Analysis (BCA; Ané et al. 2007) performs nonparametric Bayesian clustering of data subsets into groups defined by their preference for a distinct tree topology.

In our recent study involving data from 7 genes and 33 taxa in the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta), a BCA analysis using BUCKy concluded that each gene fell in its own cluster: 7 different gene trees for 7 genes. Adjustment of the prior distribution on number of clusters had no effect because no sampled tree topology was shared by any two genes, even though many splits (clades) were represented in a majority of gene trees.

We suspected that the behavior of BUCKy was due to the fact that BUCKy never considers more than one data subset at a time. Holder et al. recently published a method for accurate estimation of the phylogenetic marginal likelihood, where *phylogenetic* refers to the fact that it not only integrates over all substitution model parameters, but also over tree topologies. This makes possible Bayes Factors (BF) that compare the fit of a model when one tree topology is assumed for all subsets (the “combined” model) to the (”separate”) model where each subset is allowed to have its own tree topology (potentially different from any other subset).

We found that BF favored combining all plastid genes, all rDNA genes, and all genes, a result distinctly different than that offered by BCA. Interestingly, BF found that combining all genes was ***not*** preferable to a partition containing two subsets: a “plastid gene” (all plastid data combined) versus an “rDNA gene” (all rDNA data combined). It is probable that this mild incompatibility is masked when the “combined” model is compared to the “separate” model due to the unusually large number of parameters in the “separate” model. Bayes Factors implicitly impose a dimension penalty on models, and the 5.7-fold greater number of parameters apparently resulted in a penalty that offset any differences in goodness-of-fit.

In conclusion, using Bayes Factors to test combinability appears to be a promising new approach, and the fact that it takes into consideration both combined data and separate data may allow it to escape the extreme results of BCA when tree samples from different genes share many splits yet have no tree topologies in common.

Literature Cited

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24:412–426. (BUCKy: http://www.stat.wisc.edu/~ane/bucky/)

Fučíková, K., Lewis, P. O., and Lewis, L. (accepted). Putting *incertae sedis* taxa in their place: a proposal for ten new families and three new genera in Sphaeropleales (Chlorophyceae, Chlorophyta). *J. Phycol.*

Holder, M. T., Lewis, P. O., Swofford, D. L., and Bryant, D. (forthcoming). Variable tree topology stepping-stone marginal likelihood estimation. In: Chen, M.-H., Kuo, L., and Lewis, P. O. (eds.), *Bayesian phylogenetics: methods, algorithms, and applications*. Chapman & Hall, New York.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DEB-1036448 (GrAToL). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

UCONN
UNIVERSITY OF CONNECTICUT

NSF

GrAToL
Assembling the Green Algal Tree of Life