Testing Congruence in Phylogenomic Analysis

JESSICA W. LEIGH,[1] EDWARD SUSKO,[2] MANUELA BAUMGARTNER,[3] AND ANDREW J. ROGER[1]

[1]*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax NS, Canada B3H 1X5; E-mail: andrew.roger@dal.ca; (A.J.R.)*
[2]*Department of Mathematics and Statistics and Genome Atlantic, Dalhousie University, Halifax NS, Canada B3H 3J5*
[3]*Department für Biologie I, Botanik, Ludwig-Maximilians-Universität München, Menzingerstraße 67, D-80638 München, Germany*

Abstract.— Phylogenomic analyses of large sets of genes or proteins have the potential to revolutionize our understanding of the tree of life. However, problems arise because estimated phylogenies from individual loci often differ because of different histories, systematic bias, or stochastic error. We have developed CONCATERPILLAR, a hierarchical clustering method based on likelihood-ratio testing that identifies congruent loci for phylogenomic analysis. CONCATERPILLAR also includes a test for shared relative evolutionary rates between genes indicating whether they should be analyzed separately or by concatenation. In simulation studies, the performance of this method is excellent when a multiple comparison correction is applied. We analyzed a phylogenomic data set of 60 translational protein sequences from the major supergroups of eukaryotes and identified three congruent subsets of proteins. Analysis of the largest set indicates improved congruence relative to the full data set and produced a phylogeny with stronger support for five eukaryote supergroups including the Opisthokonts, the Plantae, the stramenopiles + Apicomplexa (chromalveolates), the Amoebozoa, and the Excavata. In contrast, the phylogeny of the second largest set indicates a close relationship between stramenopiles and red algae, to the exclusion of alveolates, suggesting gene transfer from the red algal secondary symbiont to the ancestral stramenopile host nucleus during the origin of their chloroplast. Investigating phylogenomic data sets for conflicting signals has the potential to both improve phylogenetic accuracy and inform our understanding of genome evolution. [Concatenated analysis; endosymbiotic gene transfer; hierarchical clustering; lateral gene transfer; likelihood ratio testing; maximum likelihood; phylogenetic congruence; phylogenomics; separate analysis; superkingdom; supermatrix analysis.]

The combined phylogenetic analysis of multiple genes or proteins has become popular due to the poor resolution of phylogenies based on single loci, and has been facilitated by the exponential growth of public sequence databases. In a number of recent high-profile studies, several to hundreds of genes have been combined into supermatrices to infer better-resolved phylogenies within the major taxonomic groups, including the animals (Rokas et al., 2005), plants (Philippe et al., 2005), fungi (James et al., 2006), Archaea (Brochier et al., 2005), and even the tree of life (Ciccarelli et al., 2006). Multigene or multiprotein analyses are usually predicated on the assumption that the combined genes all share the same history that is reflective of organismal relationships, and, by combining them, stochastic error in the phylogenetic estimate should be reduced (de Queiroz and Gatesy, 2007). However, gene trees and species trees do not always agree because of population-level lineage sorting (Pollard et al., 2006), hybridization (McBreen and Lockhart, 2006), gene duplication and differential loss, and lateral gene transfer (LGT), whereby genes are exchanged between lineages (Dagan and Martin, 2006; Beiko et al., 2005). In these situations, a single bifurcating tree cannot describe the disparate histories of genes under analysis. Genes may also appear to have different evolutionary histories due to inadequacy of the model used in phylogenetic inference (systematic error). Here, we define incongruence between genes as phylogenetic incompatibility, either due to truly different evolutionary history, or to systematic error. In either case, phylogenetic analysis based on the combined markers can be problematic, because there is no guarantee that the tree estimated by this approach will properly describe the history of any of the loci under consideration. Estimates of single-gene or single-protein phylogenies can also differ due to stochas-

tic error associated with the small amount of information contained in a single marker. Unfortunately, it is difficult to determine a priori whether topological differences between single-gene trees result from incongruence or from stochastic error.

Despite these difficulties, large-scale phylogenomic studies often do not explicitly deal with the issue of congruence (Rokas et al., 2005; Qiu et al., 2006; James et al., 2006) or do so in rather *ad hoc* ways (Ciccarelli et al., 2006). Nevertheless, several methods have been developed to assess congruence among markers in a phylogenetic analysis (see Planet, 2006, for a recent review). For instance, the incongruence length difference (ILD) test (Farris et al., 1995) is a parsimony-based method that compares the length of the tree inferred from the combined data set to the combined length of the trees inferred for each locus in the data set. Although initially designed to be an all-or-none congruence test, this method has been extended to allow the identification of congruent subsets of markers (Planet et al., 2003). However, there are numerous problems with the ILD test. As a parsimony-based test, it is sensitive to evolutionary conditions such as variable evolutionary rates in lineages or variation of rates across sites (Darlu and Lecointre, 2002). In addition, *P*-values from the ILD test correlate poorly with improvement in phylogenetic resolution resulting from concatenation (Barker and Lutzoni, 2002). The ILD test is therefore not particularly useful as a congruence test, particularly in a probabilistic framework.

In a maximum likelihood (ML) context, hypothesis tests such as the approximately unbiased (AU) (Shimodaira, 2002) or the Shimodaira-Hasegawa (SH) (Shimodaira and Hasegawa, 1999) test have been used to determine whether individual markers reject the tree inferred from the concatenation of all markers (Lerat

et al., 2003). This method is problematic, because the outcome is strongly dependent on topologies selected by the user. Yet another method uses principal components analysis (PCA) to cluster responses (e.g, log-likelihoods or *P*-values) of individual markers to several different candidate tree topologies (Brochier et al., 2005). Congruent markers are expected to display similar responses to different tree topologies and will therefore cluster together. However, markers with little phylogenetic signal will have similar, neutral responses to most topologies and will therefore cluster together, though this clustering due to lack of signal is not clearly equivalent to congruence (Bapteste et al., 2005; Susko et al., 2006). In addition, this method is highly sensitive to the topologies tested, and it is difficult to objectively identify clusters of congruent markers. Another likelihood-based method has been proposed that employs heat maps to cluster markers based on similar hypothesis test *P*-values for a set of tree topologies (Bapteste et al., 2005; Susko et al., 2006). Heat maps can be extremely powerful for identifying incongruence among markers, but results are largely qualitative, and they are of limited use for objectively identifying congruent subsets of markers.

Bayesian methods for the explicit estimation of multiple topologies from multigene data have recently been developed (Suchard, 2005). These methods, although promising, are computationally infeasible for the large numbers of taxa and markers typically present in comprehensive phylogenetic analyses of major taxonomic groups. Ané et al. (2006) have developed an alternative Bayesian approach whereby concordance between partitioned phylogenetic markers is estimated by a two-stage Markov chain Monte Carlo analysis. Although this method can be used for larger data sets, the posterior concordance estimates are heavily dependent on user-specified parameters of the prior distributions, limiting their usefulness in the absence of background information.

Apart from the question of whether data from separate loci should be combined at all, the choice of an appropriate method for combining these data must be considered. We restrict our focus here to two supermatrix methods of data combination. In straightforward concatenated analysis (e.g., Baldauf et al., 2000; Fitzpatrick et al., 2006), single-marker alignments are combined in a supermatrix, from which a tree is inferred. In the separate analysis method (e.g., Hasegawa et al., 1992; Bapteste et al., 2002; Simpson et al., 2006; Pupko et al., 2002), alignments themselves are not directly combined; instead, during a likelihood-based tree searching process, log-likelihoods are evaluated separately for each alignment and then summed over all alignments for a given tree. The tree that maximizes this sum is then the maximum likelihood (ML) tree. The advantage of separate analysis is that different markers that evolve under different relative lineage-specific evolutionary rates are modeled better. However, the additional parameters may not be justified, in which case the inference power is reduced by model overfitting. Hybrid methods, in which branch lengths are scaled by a marker-specific rate for each marker in a multilocus data set, have also been proposed (Yang, 1996; Pupko et al., 2002; Bevan et al., 2005).

Motivated by the shortcomings of existing methods, we have developed an application, CONCATERPILLAR, that uses hierarchical clustering and likelihood-ratio testing (LRT) to detect congruence in multigene or multiple-protein data sets. It is based on a LRT similar to that proposed by Huelsenbeck and Bull (1996) that compares the likelihood of markers forced to share a tree topology to their likelihoods when each is allowed its own tree topology. Once topological congruence is assessed, as a second stage of analysis CONCATERPILLAR uses similar methodologies to identify branch-length congruence (i.e., among topologically congruent markers), indicating which markers should be combined by concatenation, and which should have nuisance parameters separately optimized.

METHODS

Log-Likelihood Ratio Calculation

Likelihood ratios for the assessment of topological congruence are as defined in Huelsenbeck and Bull (1996). Let l_j denote a log-likelihood calculated for that data from alignment j and let $\hat{\tau}_j$, $\hat{t}_j$, and $\hat{\alpha}_j$ denote the corresponding estimated topology, edge-lengths, and shape parameter for the Γ model of rates across sites. For instance, $l_A(\hat{\tau}_{AB}, \hat{t}_A, \hat{\alpha}_A)$, denotes the log-likelihood calculated with the sites in alignment A, for the topology $\hat{\tau}_{AB}$ estimated from concatenated alignments A and B, with corresponding edge-lengths $\hat{t}_A$ and shape parameter $\hat{\alpha}_A$ estimated just using the data from A. The log-likelihood ratio is then given by:

$$\Lambda_{A,B} = l_A(\hat{\tau}_A, \hat{t}_A, \hat{\alpha}_A) + l_B(\hat{\tau}_B, \hat{t}_B, \hat{\alpha}_B)$$
$$- [l_A(\hat{\tau}_{AB}, \hat{t}_A, \hat{\alpha}_A) + l_B(\hat{\tau}_{AB}, \hat{t}_B, \hat{\alpha}_B)] \quad (1)$$

For the branch-length congruence test, the log-likelihood ratio is calculated between the likelihood of the two markers when branch lengths (and other nuisance parameters) are optimized separately and their likelihood when forced to share jointly optimized parameters (i.e., under concatenated analysis). The tree topology, $\hat{\tau}$ used for this test is inferred from the concatenation of all markers. The log-likelihood ratio is given by:

$$\Lambda_{A,B} = l_A(\hat{\tau}, \hat{t}_A, \hat{\alpha}_A) + l_B(\hat{\tau}, \hat{t}_B, \hat{\alpha}_B)$$
$$- [l_A(\hat{\tau}, \hat{t}_{AB}, \hat{\alpha}_{AB}) + l_B(\hat{\tau}, \hat{t}_{AB}, \hat{\alpha}_{AB})] \quad (2)$$

For more complex evolutionary models than those currently implemented in CONCATERPILLAR, additional parameters would necessarily be included in Equations 1 and 2.

Inference of Phylogenies and Likelihood Calculation

Given that trees must be inferred for all markers, as well as all pairs of markers (for n markers, a total of $\frac{1}{2}(n^2 + n)$ trees are estimated), a reasonably quick

inference method was required. Consequently, phylogenetic trees are inferred with PHYML (Guindon and Gascuel, 2003). Likelihoods for trees produced from concatenated pairs of markers are then assessed for relevant single markers (e.g., for the tree inferred from concatenated markers A and B, $\hat{T}_{AB}$, likelihoods $l^A_{\hat{T}_{AB}}$ and $l^B_{\hat{T}_{AB}}$ are calculated). For this likelihood calculation, TREE-PUZZLE (Schmidt et al., 2002) is used. For both TREE-PUZZLE and PHYML, the substitution model is selected by the user. Rates across sites is modeled by a four-category discretized Γ distribution. During tree estimation, the shape parameter is optimized by PHYML for every tree inferred. For the TREE-PUZZLE–based likelihood calculation, the shape parameter estimated by PHYML for an individual data set (i.e., a single marker or set of congruent markers) is used to evaluate the likelihood of the single data set under any tree considered.

Assessment of Significance

In the test for topological congruence, after likelihood ratios are determined for all pairs of markers, the pair with the smallest likelihood ratio (i.e., the pair least likely to reject congruence) is chosen, and a P-value (the probability of observing the likelihood ratio if the two markers were congruent) is determined. Due to the discrete nature of tree-topologies, χ^2 distributions cannot be used to calculate the P-value. Instead, a bootstrapping method is used. Nonparametric bootstrapping was chosen over parametric bootstrapping in order to avoid effects of model misspecification. For half of the bootstrap replicates, columns are drawn from one of the two aligned markers, while they are drawn from the other marker for the remaining replicates. Assuming all sites within a single marker are congruent, this technique ensures that resampled alignments are topologically congruent for the null distribution.

Although a similar procedure is used in the assessment of significance for the branch-length congruence test, no bootstrapping is required. Under the null hypothesis of congruence, twice the likelihood ratio used in this test is χ^2 distributed with degrees of freedom equal to the difference in number of parameters between the two models compared. In this case, there are $2n - 2$ additional parameters when each marker is allowed its own branch lengths and shape parameter for Γ distributed rates across sites.

For either test, if the P-value is larger than the user-defined cutoff (α level), congruence is not rejected, and the pair is combined. The test then continues, treating this pair as a single marker. If, however, the P-value falls below the α level, congruence is rejected and the test ends (Fig. 1).
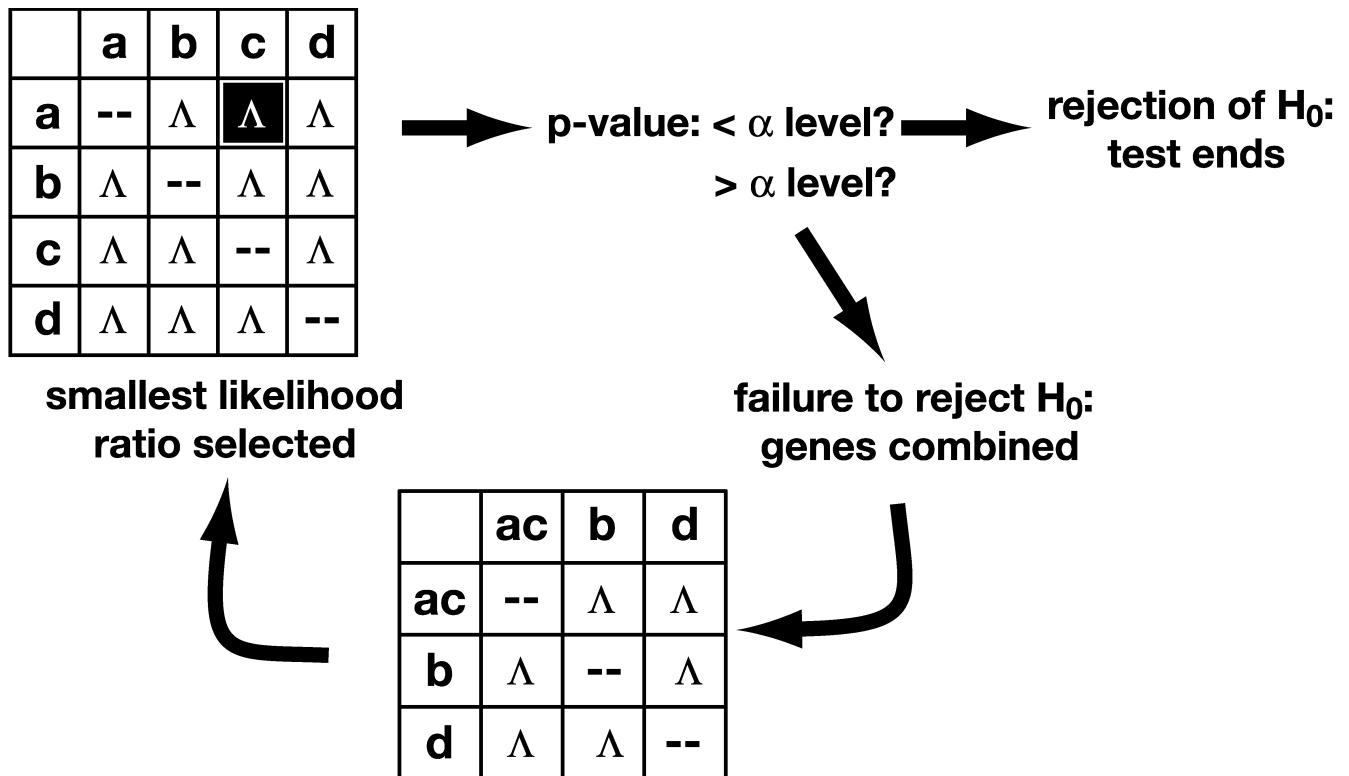

FIGURE 1. General overview of the CONCATERPILLAR algorithm. For both congruence tests, CONCATERPILLAR follows a similar algorithm. For all pairs of markers, a log-likelihood ratio is calculated. A p-value is estimated for the pair with the smallest ratio (i.e., the most congruent pair). If the p-value falls below the user-defined α level, congruence is rejected, and the test ends. Otherwise, the markers are combined, and the test continues, with the two markers treated as one thereafter.

Multiple Comparison Corrections

The methodology used in CONCATERPILLAR results in two opposing multiple-testing problems. First, the repetition of the likelihood-ratio test over the levels of the hierarchy (Fig. 1) results in an increase in the probability of type I error (false rejection of congruence) at some level of the hierarchy as the number of levels increases. Secondly, the probability of type I error decreases with the number of likelihood-ratios compared at a given level of the hierarchy (i.e., the number of phylogenetic markers or sets of combined markers). Treating individual tests as independent, the two errors can be accommodated by adjusting the α level. Congruence is rejected when likelihood-ratio test P-values are less than the adjusted α level. Let α_u denote the user-defined α level (e.g., 0.05), k the number of levels in the hierarchy, and c the number of independent comparisons made at a given level of the hierarchy. The adjusted α level, α_c, is then:

$$\alpha_c = \left[1 - (1 - \alpha_u)^{\frac{1}{k}} \right]^{\frac{1}{c}} \qquad (3)$$

Under the hypothesis that all genes are congruent (H_0), k is one less than the total number of markers tested, and c varies throughout the hierarchy, and is approximated here by half the number of markers (or clusters of markers), n, at any given level of the hierarchy. This is an approximation because, even though there are $\left(\binom{n}{2} \right)$ comparisons made at each level of the hierarchy, only (the largest integer not exceeding) $\frac{n}{2}$ of these, corresponding to non-overlapping concatenations, are truly independent. Thus, the correction under the null is given by:

$$\alpha_c = \left[1 - (1 - \alpha_u)^{\frac{1}{k}} \right]^{\lfloor \frac{n}{2} \rfloor^{-1}} \qquad (4)$$

However, our simulation analyses indicated that this correction may be too stringent. In cases where H_0 is not true (that is, at least some markers are incongruent), many of the comparisons made at a given hierarchical level will correspond to the alternative hypothesis, for which P-values are expected to be smaller than those predicted for H_0. Consequently, a corrected α level based on H_0 will be larger than necessary. We have also investigated the performance of some alternative corrections. First, we estimated the number of clusters using the uncorrected, user-defined α level, α_u. We then applied Equation 3, defining k as the predicted number of levels and c as the sum of half the number of markers (n_i) in each predicted cluster (c varies throughout the hierarchy). This within-cluster correction then becomes:

$$\alpha_c = \left[1 - (1 - \alpha_u)^{\frac{1}{k}} \right]^{\left(\sum_{i=1}^{C} \lfloor \frac{n_i}{2} \rfloor \right)^{-1}} \qquad (5)$$

where C is the number of clusters. In this case, the correction takes into account only within-cluster comparisons, those comparisons for which H_0 is true. For highly congruent data sets, the number of clusters will be smaller, and the number of within-cluster comparisons will increase. As a result, α_c will be increased. This correction is logical because P-values in more highly congruent data sets are likely to be higher, maintaining the meaning of α_u as the probability of type I error. On the other hand, it might be more appropriate to use an α_c that favors combination of markers when data are largely congruent, and penalizes clustering when less congruence is predicted. Consequently, we have also examined the performance of a correction formula that takes into account only the predicted number of levels in the test hierarchy. The formula for this hierarchy-only correction is given by:

$$\alpha_c = 1 - (1 - \alpha_u)^{\frac{1}{k}} \qquad (6)$$

Simulations

The performance of both the topological and branch-length congruence tests was evaluated using amino acid sequences simulated using Seq-Gen (Rambaut and Grassly, 1997) under various evolutionary scenarios. In all cases, proteins were simulated under WAG+Γ. JTT+Γ was used in phylogenetic inference and likelihood calculation in CONCATERPILLAR, in order to simulate slight model misspecification. For the topological congruence test, 10 alignments of 10 sequences were simulated either all under the same topology (but with different branch lengths), under 10 different topologies (a different topology for each alignment), under 9 different topologies (two alignments shared a topology, each of the eight others was simulated under its own topology), or under 3 different topologies (five alignments shared one topology, three shared a second, and two shared a third topology). It should be noted that, due to the additional time required for multiple simulations with CONCATERPILLAR, the number of alignments used in these simulations is considerably smaller than might be included in a typical phylogenomic analysis. The topologies for these simulations were inferred from single-protein and concatenated alignments chosen from a set of 60 translational proteins described below (see also Supplemental Materials, available at http://www.systbio.org). For each of these scenarios, 100 simulations were performed. CONCATERPILLAR was used to identify topologically congruent sets for each simulation, using an uncorrected α level of 0.05, as well as corrections of this value given in Equations 4 to 6.

For the branch-length test, we analyzed 100 simulated 10-protein data sets generated from a single 10-sequence topology that was chosen by concatenating 10 alignments from among 60 eukaryotic translational proteins (described below). The alignments were simulated with either the same branch lengths and α parameter; different branch lengths (and α parameter) for all alignments; shared α and branch lengths for two alignments, but different parameters for the eight other proteins; or three sets of branch lengths and α parameters (one set of parameters shared for five alignments, another set for three alignments, and a third set for the remaining two alignments). Branch lengths and α parameters used for these

simulations were all chosen from maximum likelihood estimates for single or concatenated alignments from among the 60 eukaryotic translational proteins described below (see also Supplemental Materials). Once again, branch-length congruent sets were identified using CONCATERPILLAR with an uncorrected α level of 0.05 and the three multiple-comparison corrections of this value.

Receiver operating characteristic (ROC) curves (Zweig and Campbell, 1993) were plotted separately for the branch-length and topological congruence tests in order to evaluate the performance of the tests using each of the multiple comparison correction formulas. For each correction of α levels between 0 and 1, with increments of 0.01, the proportion of pairs of congruent loci correctly assigned to the same cluster was plotted against the proportion of incongruent loci incorrectly assigned to the same cluster.

Global Eukaryotic Phylogeny

Alignments of 60 ribosomal proteins from Bapteste et al. (2002) were kindly provided by Hervé Philippe. The taxonomic representation in these alignments was enhanced and missing data were filled in by manually adding sequences from the GenBank database using standard searching methods. In addition, the sequences for these 60 proteins from *Naegleria gruberi* were obtained from an expressed sequence tag (EST) project that will be described elsewhere (Sjögren, Gill, and Roger, unpublished data). Alignments were visually inspected and ambiguously aligned regions were excluded from further analysis. The final data set had 60 proteins, 22 species, and 9532 total sites. All data sets were deposited in TreeBASE under accession number SN3537.

The 60 alignments were analyzed for topological congruence using CONCATERPILLAR with an initial α level of 0.05, and the total number of levels of the hierarchy was predicted via a single round of uncorrected analysis as described above. The α level was then corrected based on the predicted number of test iterations using Equation 6, as this method performed best overall in the simulation analyses. Phylogenetic analysis in CONCATERPILLAR used JTT+Γ4, with the shape parameter estimated from the data.

For the set of 60 proteins, as well as each topologically congruent set, proteins were concatenated and a tree was inferred using IQPNNI (Vinh le and Von Haeseler, 2004) with WAG+Γ4, and bootstrap support was determined from 100 replicates. Additional bootstrap support values (BSJK60) for the set of all 60 proteins were determined using a combination of jackknife and bootstrap resampling in order to produce support values that would be more easily comparable to those obtained from the largest congruent set of proteins. In this method, 6243 columns (the number of sites in the larges topologically congruent set) were chosen at random from among the 9532 sites in the concatenated 60-protein alignment. These sites were then resampled with replacement to produce a bootstrapped alignment with 6243 positions, from which a tree was inferred. This jackknife + bootstrap process was repeated 100 times.

Each set of congruent proteins was analyzed with CONCATERPILLAR's branch-length congruence test in order to determine which proteins should be analyzed separately (again, an initial α level of 0.05 was corrected based on the predicted number of test levels). For the largest congruent set, those proteins found to have congruent branch lengths were concatenated, and the resulting set of proteins and concatenated sets of proteins were analyzed separately using an exhaustive search strategy with constraints on certain nodes of the tree. Opisthokonta, Sarcocystidae + *Plasmodium*, *Chlamydomonas* + land plants, Amoebozoa, and Excavata were constrained, and all resulting 945 trees were evaluated by separately calculating the likelihood for each branch-length congruent set using TREE-PUZZLE, and log-likelihoods were summed over all sets. RELL bootstrap support was determined by resampling (with replacement) site wise likelihoods individually from each protein and choosing the best tree for each of 10,000 replicates. For comparison, RELL support was also determined from the concatenation of all proteins in this set, using the same set of 945 trees.

RESULTS AND DISCUSSION

CONCATERPILLAR *Accurately Identifies Incongruence*

We have developed an application, CONCATERPILLAR (available from http://www.rogerlab.biochem.dal.ca/Software/Software.htm), in which we have implemented methods to test for two kinds of hypotheses in supermatrix analysis. The first is the null hypothesis (H_0) that the phylogenies of markers in the supermatrix are congruent. If we cannot reject congruence for a set of markers, the second hypothesis to test is whether or not the markers to be combined have significantly different evolutionary dynamics (branch lengths and rates-across-sites parameters); that is, whether they should be concatenated or subjected to separate analysis.

In order to determine the accuracy with which CONCATERPILLAR identifies topological congruence, we evaluated its performance with data simulated under four scenarios: A, complete congruence; B, three congruent sets; C, two congruent and eight incongruent proteins; and D, complete incongruence. Table 1 shows the results from the various α level corrections and test scenarios as the frequency with which pairs of proteins were correctly or incorrectly identified as either congruent or incongruent. The performance of the corrections depended heavily on the degree of congruence amongst the proteins. In highly congruent scenarios (three sets or complete congruence), correcting under H_0 or for the number of within-cluster comparisons resulted in considerably poorer performance than when the hierarchy-only correction was applied; the use of an uncorrected α level also resulted in poor performance when all proteins were congruent. When all proteins were incongruent, all the corrections did well. The case where there was

TABLE 1. Performance of the topological congruence test under A: complete congruence, B: three congruent sets; C: only two congruent proteins; and D: complete incongruence.

True classification	CONCATERPILLAR prediction							
	No correction		Under H_0		Within-cluster		Hierarchy only	
	T	NT	T	NT	T	NT	T	NT
A								
T	0.885	0.115	0.862	0.138	0.878	0.122	0.986	0.014
NT	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
B								
T	0.959	0.041	0.799	0.201	0.899	0.101	0.994	0.006
NT	0.004	0.996	0.010	0.990	0.012	0.988	0.012	0.988
C								
T	0.960	0.040	0.790	0.210	0.920	0.080	0.980	0.020
NT	0.058	0.942	0.017	0.983	0.079	0.921	0.090	0.910
D								
T	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
NT	0.037	0.963	0.007	0.993	0.047	0.953	0.054	0.946

T = proteins clustered together; NT = proteins did not cluster together.

a single pair of congruent proteins with all others incongruent was the most difficult to correctly recover, and the correction under the null did particularly poorly in this case. By contrast, the hierarchy-only correction did well under all of the various conditions. We investigated the performance of the corrections further by plotting ROC curves for all four corrections for all four simulation conditions combined (Fig. 2a). The ROC curves indicate that all of the methods do reasonably well, with the hierarchy-only correction showing the best overall performance and the within-cluster correction the poorest.

A similar set of simulations was used to evaluate the effectiveness of the branch-length congruence test. In this case, sets of 10 proteins were all simulated under the same topology but with either the same or different sets of branch lengths. Again, there were four sets of simulations: A, all proteins were simulated with the same branch lengths; B, three sets of branch lengths; C, only two proteins shared branch lengths; and D, all proteins were simulated with different branch lengths. Once again, the hierarchy correction outperformed other formulas (Table 2, Fig. 2b).

Both the topology and branch-length tests were able to accurately identify congruence when the hierarchy correction was applied. Surprisingly, type I error was much higher in the branch-length congruence test than in the topological congruence test, regardless of the correction formula used. The source of this discrepancy is unclear but may have to do with easier discrimination between discrete objects like topologies, in comparison to continuous objects like branch lengths that can differ but be very similar. In any case, increased type I error will bias the branch-length test towards rejecting congruence, resulting in the separate analysis of some proteins that should be concatenated, and increasing the variance of the resulting phylogenetic estimate. However, this increase in random error seems acceptable when weighed against the potential for systematic error incurred by falsely concatenating proteins with different branch length sets (e.g., Kolaczkowski and Thornton, 2004).

Exclusion of Incongruent Markers Improves Phylogenetic Resolution for Eukaryotic Supergroups

To test CONCATERPILLAR on a real data set, we applied it to estimating superkingdom-level relationships amongst eukaryotes with 60 alignments of translational components including ribosomal proteins, initiation factors, and elongation factors (Supplemental Materials). CONCATERPILLAR's topological congruence test was used to identify congruent sets of proteins using an initial α level (α_u) of 0.05, which was then corrected based on the predicted number of hierarchical levels (Equation 6), because this correction clearly performed best with simulated data. Applying the uncorrected α level results in a prediction of 53 levels. Substituting $k = 53$ into Equation 6, α_c becomes 9.67×10^{-4}, which results in rejection at the 58th level. Reiteration of the correction formula with $k = 58$ produces an α_c of 8.84×10^{-4}, which results once again in rejection of H_0 at level 58 (i.e., α_c has converged). These sets remained stable with an α level of 0.01 with the same multiple comparisons correction. From these corrections, three mutually incongruent sets of proteins were identified, containing 35, 15, and 10 proteins, respectively. ML phylogenies inferred from the concatenation of all of the proteins as well as in each of the three sets are shown in Fig. 3.

The topology based on all 60 proteins (Fig. 3a) showed five superkingdom-level groups of eukaryotes that have been proposed based on a variety of other data (Keeling et al., 2005; Simpson and Roger, 2004), including the Plantae, the Chromalveolates, the Excavata, the Amoebozoa, and the Opisthokonta. Interestingly, however, the bootstrap support for these groupings in some cases is relatively weak (e.g., 53% for Chromalveolata and 56% for Plantae) despite the relatively large size of this data set. Not surprisingly, the topology inferred from the largest congruent set (35 proteins; Fig. 3b) is topologically similar to the 60-protein topology, differing only in the positions of *Naegleria* and of *Caenorhabditis*. More interestingly, the bootstrap support values for many groups changed substantially, with bootstrap support values obtained from the set of 35 proteins (BS35) generally increasing relative to boostrap values obtained from the 60-protein data set (BS60). The support for Plantae increased from 56% to 91%, support for stramenopiles + Apicomplexa (chromalveolates) increased from 53% to 91%, support for Plantae + Chromalveolata increased from 85% to 99%, and support for Excavata + Amoebozoa increased from 79% to 97%.

These increases in bootstrap support are observed despite a marked reduction in the number of amino acid characters in the 35-protein set (6243, compared to 9532 sites in the 60-protein set). In order to compare support values from an equivalent number of positions, a jackknife + bootstrap method was used to obtain additional support values (BSJK60) from the 60-protein set. The difference between BSJK60 and BS35 values was even greater than when BS35 and BS60 values were compared: support for Plantae dropped to 49%, for chromalveolates to 40%, 77% for Plantae + Chromalveolata, and 69% for
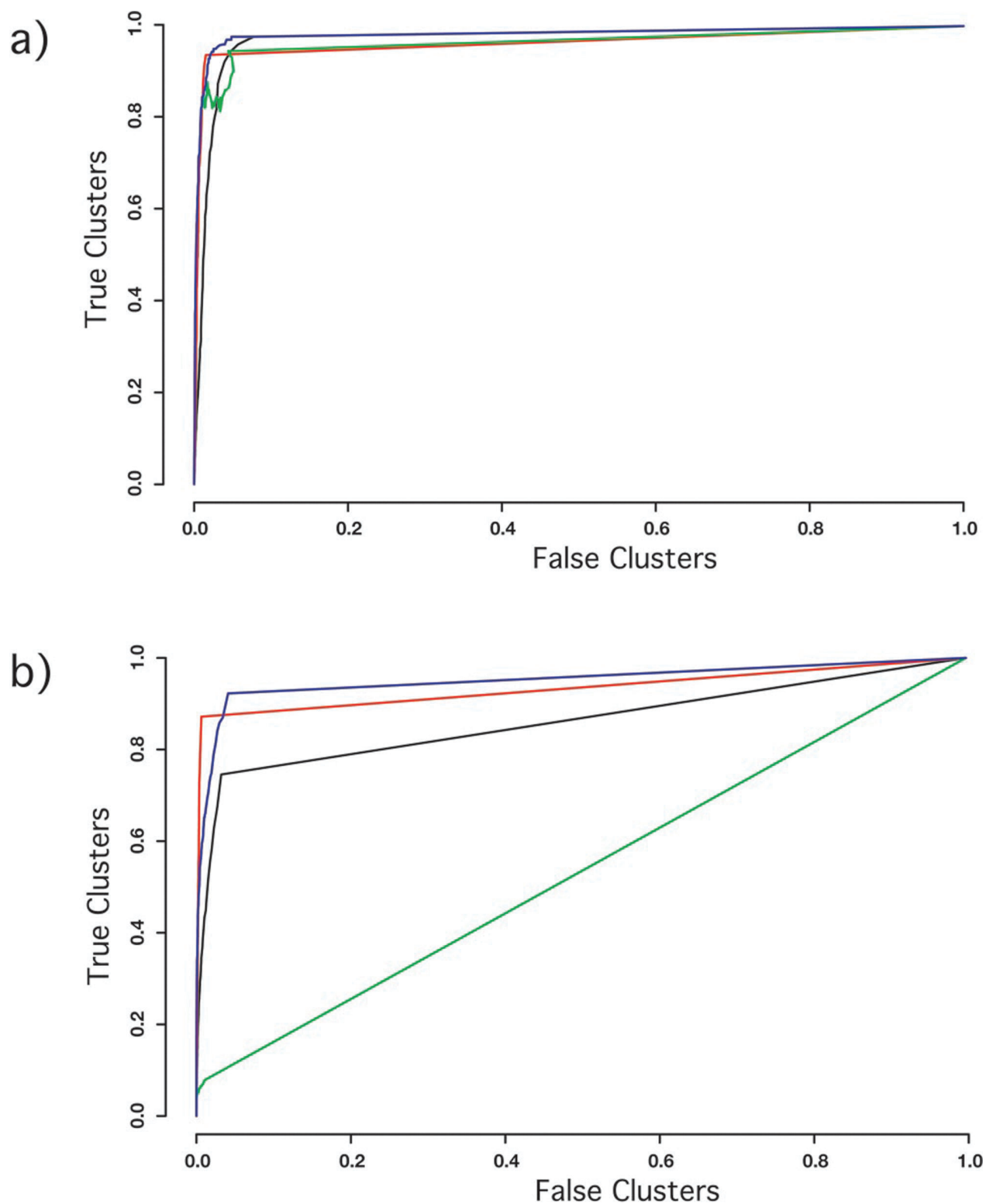
FIGURE 2.   ROC curves from simulation results with (a) topological congruence test and (b) branch-length congruence test. The best possible performance of a classifier on an ROC diagram is indicated by a curve that is in the top left-hand corner (high true-positive rate and low false-positive rate), with the curve of a random classifier falling on the $x = y$ line. For user-defined α levels between 0 and 1, CONCATERPILLAR was used to identify congruence among simulated data with varying levels of congruence, using different correction formulas of the α level: black, no correction; red, correction under H_0; green, within-cluster correction; blue, hierarchy correction. For both the topological congruence test (a) and branch-length congruence test (b), the number of proteins that were correctly clustered together was plotted against the number of proteins that were falsely clustered together for each α level.

TABLE 2. Performance of the branch-length congruence test under A: complete congruence; B: three congruent sets; C: only two congruent proteins; and D: complete incongruence.

| True classification | CONCATERPILLAR prediction | | | | | | | |
| | No correction | | Under H_0 | | Within-cluster | | Hierarchy only | |
	T	NT	T	NT	T	NT	T	NT
A								
T	0.481	0.519	0.752	0.248	0.044	0.956	0.770	0.230
N	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
B								
T	0.781	0.219	0.681	0.319	0.114	0.886	0.942	0.058
NT	0.002	0.998	0.002	0.998	0.000	1.000	0.002	0.998
C								
T	0.810	0.190	0.470	0.530	0.690	0.310	0.840	0.160
NT	0.025	0.975	0.005	0.995	0.008	0.992	0.036	0.964
D								
T	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
NT	0.022	0.978	0.004	0.996	0.015	0.985	0.029	0.971

T = proteins clustered together; NT = proteins did not cluster together.

Excavata + Amoebozoa. For a few relationships, a notable increase in support was observed from BSJK60 to BS35, where no significant difference had been noted when BS35 had been compared to BS60: the monophyly of excavates (BSJK60 = 81%; BS60 = 95%; and BS35 = 98%) and of Amoebozoa (BSJK60 = 75%; BS60 = 90%; and BS35 = 95%). This result strongly suggests increased congruence in the smaller set of proteins. However, it is worth noting that, although the set of 35 proteins is the largest, its phylogeny does not necessarily represent the true organismal phylogeny: it is possible that these proteins simply share similar features that result in phylogenetic artefact (e.g., long-branch attraction). For example, the grouping of Amoebozoa and Excavata is not supported by other data that we are aware of and conflicts with the proposed rooting of the eukaryote tree using gene fusion and gene family data between so-called "bikont" groups (Chromalveolates, Excavata, Plantae) and "unikont" groups (Opisthokonta and Amoebozoa; see Richards and Cavalier-Smith, 2005, and references therein).

Although CONCATERPILLAR's performance is excellent, as demonstrated by our simulations, a potential problem arises from the multiple comparison correction, which results in reduction of the α level; in this case, the corrected α became 8.84×10^{-4}, which resulted in rejection of the null hypothesis with a P-value of 0. However, this P-value was determined from one hundred bootstrap replicates; due to lack of precision, the very small α level may result in false rejection of the null hypothesis. Although increasing the number of bootstrap replicates would improve the precision of the estimated P-value, it would also result in a drastic increase in the run-time of CONCATERPILLAR. Instead, we fitted the shape of the distribution of bootstrap log-likelihood ratios to an appropriate statistical distribution and then estimated the P-value from this distribution. Of several tested, a Weibull distribution fits the bootstrap values best (Supplemental Materials). For our set of 60 proteins, modeling the bootstraps with a Weibull distribution resulted in the identification of the same three clusters, with a P-value of 0.

Separate Analysis Lowers Bootstrap Support

For the topologically congruent set of 35 proteins, CONCATERPILLAR's branch-length congruence test was used to identify proteins that could be concatenated. An initial α level of 0.05 was corrected based on the predicted number of levels in the test hierarchy. CONCATERPILLAR identified 23 sets of proteins that should be analyzed separately, of which 12 contained only one protein, 10 contained two proteins, and 1 contained three proteins. These 23 sets were analyzed separately using an exhaustive tree search strategy, with constraints on certain nodes. For those nodes that were not constrained, bootstrap support from resampling of estimated log-likelihoods (RELL; Kishino et al., 1990) is shown in Fig. 3b. RELL bootstrap support values are also shown for the concatenation of these 23 sets.

Compared to the concatenated analysis, bootstrap support for key branches decreased somewhat when separate analysis was used, but in many cases the decrease was small. Even in the absence of model misspecification, this observation is expected because of the increase in variance due to the increased number of parameters associated with separate analysis relative to concatenated analysis. Interestingly, the largest decrease in bootstrap support was observed for Amoebozoa + Excavata, which dropped from 91% to 80%. Because this grouping is probably incorrect, it seems that, although both analysis methods are affected by the same systematic error, this error is less prominent under separate analysis. Interestingly in this case, the alternative hypothesis, Amoebozoa + Opisthokonta, which is consistent with the unikont/bikont rooting of eukaryotes, increases in bootstrap support from 9% in concatenated analysis to 20%. Thus at least some, but not all, of the support for the Amoebozoa + Excavata grouping can be accounted for by model misspecification from concatenating proteins that should be separately analyzed. Other forms of model misspecification, such as amino acid compositional heterogeneity (Foster and Hickey, 1999) and site-specific substitution processes (Lartillot and Philippe, 2004), may contribute to the support for this grouping, but a thorough investigation of the causes of phylogenetic artifacts in this data set is beyond the scope of this study.

Endosymbiotic Gene Replacement of Several Ribosomal Proteins

Although there are many unusual phylogenetic relationships that appear in trees inferred from the smaller congruent sets of proteins (particularly the 10-protein set), they are generally either poorly supported by bootstrap analysis, or appear to result from long-branch attraction. However, one particularly interesting result with biological implications comes from analysis of the second-largest congruent set. In this set of 15 proteins, the stramenopiles group with the rhodophyte alga (*Porphyra*) to the exclusion of the Apicomplexa with a bootstrap support value of 88%. Stramenopiles + Apicomplexa were supported by a value of 91% in the tree based on the set of 35 proteins (86% under
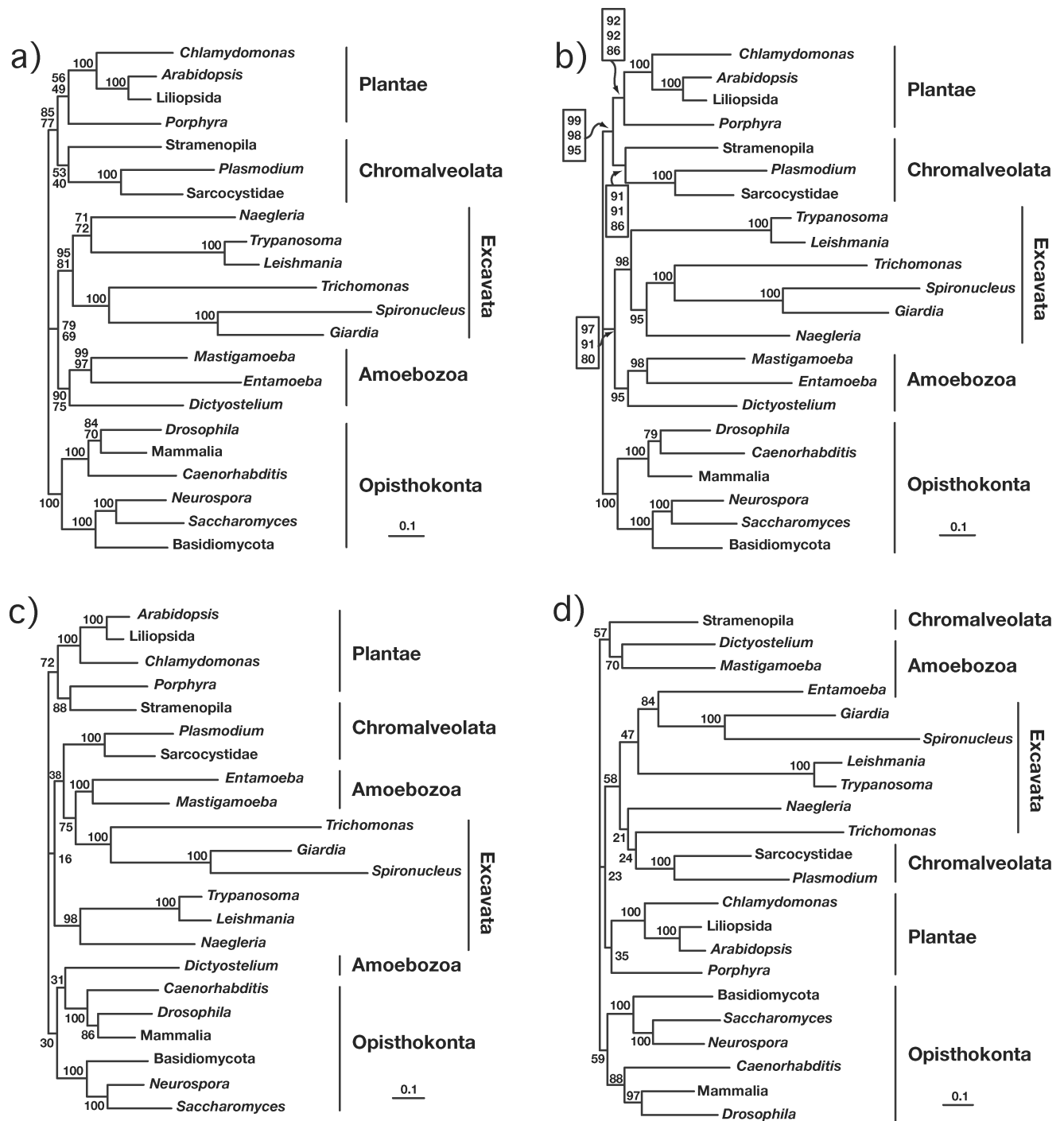
FIGURE 3. Maximum likelihood phylogenies inferred from 60 eukaryotic translational proteins and congruent subsets. (a) Tree for the concatenated set of all 60 proteins. Bootstrap support from the entire set of 60 proteins (top value) and support values from jackknife resampling of 6243 positions, followed by bootstrap resampling of these sites (bottom value) are indicated. In cases where bootstrap and jackknife + bootstrap support values were identical, only one value is shown. (b) Tree from the congruent subset of 35 proteins. Bootstrap support using three different methods is indicated: top value, support from IQPNNI-based analysis of bootstrap samples based on the 35 proteins, concatenated; middle value, support from bootstrap analysis using the RELL technique of the 35 proteins, concatenated; lower value, support from RELL-based bootstrap analysis of the 35 proteins using separate analysis of branch-length congruent subsets (23 sets). RELL bootstrap support values are only given for nodes that were not constrained in the exhaustive search. (c) Tree from the congruent set of 15 proteins, concatenated. Bootstrap support is indicated. (d) Tree from the congruent set of 10 proteins, concatenated. Bootstrap support is indicated.

separate analysis), which is in accordance with the Chromalveolate hypothesis (Cavalier-Smith, 1999), whereby stramenopiles, alveolates (including apicomplexa), haptophytes, and cryptomonads form a monophyletic group whose common ancestor harbored a secondary plastid of red algal origin. The signal uniting the red algae and stramenopiles to the exclusion of the apicomplexa suggests that the genes encoding these 15 proteins may have been transferred from the red algal endosymbiont prior to the loss of its nucleus (the nucleomorph) but subsequent to the divergence of alveolates and stramenopiles. Without the use of CONCATERPILLAR, detection of these endosymbiotic gene transfers would be very difficult, because phylogenies inferred from the individual proteins are too poorly resolved (bootstrap support <50% for key branches, data not shown) to confirm the relationship between rhodophytes and stramenopiles (nor do they confirm the monophyly of either Plantae or chromalveolates). Clustering of genes with CONCATERPILLAR, however, allows the common signal supporting the relationship between rhodophytes and stramenopiles to emerge.

Limitations of CONCATERPILLAR and Alternative Methods

We have demonstrated the utility of CONCATERPILLAR in assessing congruence in large, multilocus data sets. However, it must be noted that this method scales poorly with very large numbers of loci, because the required evaluation of trees for all pairs of genes results in a computational complexity of $O(n^2)$, where n is the number of markers in the data set. Consequently, analysis of data sets with upwards of 150 markers and 30 to 40 taxa will be impractical without access to significant computational resources. For this reason, the simulations presented here included only ten alignments of ten sequences each, much smaller data sets than would normally be included in a truly phylogenomic analysis. The hierarchical correction for the α level (Equation 6) was chosen based on its performance with these simulations. As computational power increases, the performance of CONCATERPILLAR, particularly with respect to the choice of an appropriate α level correction, should be assessed with larger data sets.

In addition, care must be taken in interpreting results obtained with CONCATERPILLAR. Although it is tempting to assume that a phylogeny inferred from the largest set of congruent markers is the species phylogeny, there are many scenarios where this will not be the case. For example, coalescence theory predicts situations in which gene trees that do not reflect the species phylogeny are actually more likely than those that do, for certain combinations of branch lengths (Kubatko and Degnan, 2007). Similarly, one can envision scenarios involving LGT, paralogy, and systematic error in which the species tree is not recovered from the largest set of congruent markers.

CONCATERPILLAR uses a hierarchical likelihood-ratio testing framework to assess congruence among markers. Alternatives to the hierarchical clustering method can be imagined, such as consideration of all possible partitioning schemes. Because of the computational challenges of model fitting in phylogenetic contexts, considering all possible partitions is not at all feasible. The hierarchical nature of aggregation provides an appropriate compromise as information about pairwise comparisons obtained at one level of the hierarchy can be used at other levels. A "top-down" alternative, in which the set of all markers is iteratively split into subsets, might be reasonable, but searching over all possible partitions of a concatenated data set into two subsets would be far more computationally intensive than the "bottom-up" approach we have implemented.

Alternatives to the likelihood ratio used as a predictor of congruence can also be imagined. Popular choices are the Akaike or Bayesian information criteria (AIC or BIC; Akaike, 1974; Schwarz, 1978). For the branch-length congruence problem, these options are reasonable alternatives to the likelihood ratio. However, this approach will not work for topological congruence. In AIC, richer models with more parameters are penalized by subtracting the number of parameters from the log likelihood. An appropriate penalty for the increase in model richness from the introduction of separate topologies is not straightforward.

Finally, although we have refrained from combining the incongruent sets of proteins identified in this study, there are methods of combining data that can accommodate incongruence in an appropriate way. For instance, a supernetwork inferred from congruent subsets of markers could represent their conflicting histories (Huson and Bryant, 2006). Alternatively, likelihood-based mixture models can be conceived that would allow simultaneous estimation of multiple topologies for multiple gene sets with additional parameters that control the numbers of topologies estimated and their associated weights. Such methods will be complex to implement and will be extremely computationally burdensome but are definitely worth pursuing in future.

CONCLUSION

In this data-driven age of research, the analysis of large, multilocus data sets has become popular in phylogenetics. We have developed CONCATERPILLAR, an application that assesses both topological and branch length congruence in such data sets by means of hierarchical likelihood-ratio tests. Our results with simulated data demonstrate that our method is highly effective when the data have evolved according to different underlying trees, representing scenarios of LGT, paralogy, or lineage sorting. Similarly, the test for branch-length congruence effectively recovered clusters of markers simulated according to trees with the same branch lengths.

Our results for CONCATERPILLAR applied to the 60 translational proteins are particularly interesting. As these proteins are essential components of the translation apparatus of eukaryotes, we did not expect to find evidence for different evolutionary histories. However, our results indicate that there are three incongruent sets of proteins. Although these sets do not necessarily represent different histories, nor is any one of them necessarily

representative of the true evolutionary history of these taxa, the largest set of congruent proteins recovers strong bootstrap support for five eukaryotic supergroups that have been proposed on other biological grounds, so it is likely that much of the structure of this tree is truly reflective of historical relationships among these groups of eukaryotes. Furthermore, given our knowledge of the role of secondary endosymbiosis in the evolution of chromalveolates, the position of stramenopiles in the tree inferred from the set of 15 proteins is also likely to be biologically meaningful. In contrast, there is little in the 10-protein tree that can be explained by prior biological knowledge (and indeed, very few of the groups in this tree that do not appear in the other trees are even reasonably well-supported). Consequently, the phylogeny of this set of proteins is almost certainly affected by systematic biases. As with all large-scale data analyses, background biological knowledge and reasoning are required for reasonable interpretation of the results.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. IEEE T. Automat. Contr. 19:716–723.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2006. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, entamoeba, and mastigamoeba. Proc. Natl. Acad. Sci. USA 99:1414–1419.

Bapteste, E., E. Susko, J. Leigh, D. Macleod, R. L. Charlebois, and W. F. Doolittle. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. 5:33.

Barker, F. K. and F. M. Lutzoni. 2002 The utility of the incongruence length difference test. Syst. Biol. 51:625–637.

Beiko, R. G., T. J. Harlow, and M. A. Ragan. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. USA 102:14332–14337.

Bevan, R. B., B. F. Lang, and D. Bryant. 2005 Calculating the evolutionary rates of different genes: A fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Syst. Biol. 54:900–915.

Brochier, C., E. Bapteste, D. Moreira, and H. Philippe. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18:1–5.

Brochier, C., P. Forterre, and S. Gribaldo. 2005. An emerging phylogenetic core of archaea: Phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol. Biol. 5:36.

Cavalier-Smith, T., 1999. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. 46:347–366.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Dagan, T., and W. Martin. 2006. The tree of one percent. Genome Biol. 7:118.

Darlu, P., and G. Lecointre. 2002. When does the incongruence length difference test fail? Mol. Biol. Evol. 19:432–437.

de Queiroz, A., and J. Gatesy. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.

Farris, J. S., M. Kallersjo, A. G. Kluge, and C. Bult. 1995. Constructing a significance test for incongruence. Syst. Biol. 44:570–572.

Fitzpatrick, D. A., C. J. Creevey, and J. O. McInerney. 2006. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the rickettsiales. Mol. Biol. Evol. 23:74–85.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both dna-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284–290.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hasegawa, M., Y. Cao, J. Adachi, and T. Yano. 1992. Rodent polyphyly? Nature 355:595.

Huelsenbeck, J. P., and J. J. Bull. 1996 A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.

Huson, D. H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

James, T. Y., F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, and et al. 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443:818–822.

Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray. 2005 The tree of eukaryotes. Trends Ecol. Evol. 20:670–676.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lartillot, N., and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-proteobacteria. PLoS Biol. 1:E19.

McBreen, K., and P. J. Lockhart. 2006. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci. 11:398–404.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol. Biol. Evol. 22:1246–1253.

Planet, P. J. 2006. Tree disagreement: Measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39:86–102.

Planet, P. J., S. C. Kachlany, D. H. Fine, R. DeSalle, and D. H. Figurski. 2003. The widespread colonization island of actinobacillus actinomycetemcomitans. Nat. Genet. 34:193–198.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006 Widespread discordance of gene trees with species tree in drosophila: Evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.

Qiu, Y. L., L. Li, B. Wang, Z. Chen, V. Knoop, M. Groth-Malonek, O. Dombrovska, J. Lee, L. Kent, J. Rest, G. F. Estabrook, et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proc. Natl. Acad. Sci. USA 103:15511–15516.

Rambaut, A., and N. C. Grassly. 1997. Seq-gen: An application for the Monte Carlo simulation of dna sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Richards, T. A., and T. Cavalier-Smith. 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436:1113–1118.

Rokas, A., D. Kruger, and S. B. Carroll. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. Tree-puzzle: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Simpson, A. G., Y. Inagaki, and A. J. Roger. 2006 Comprehensive multi-gene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol. Biol. Evol. 23:615–625.

Simpson, A. G., and A. J. Roger. 2004. The real "kingdoms" of eukaryotes. Curr. Biol. 14:R693–R696.

Suchard, M. A. 2005. Stochastic models for horizontal gene transfer: Taking a random walk through tree space. Genetics 170:419–431.

Susko, E., J. Leigh, W. F. Doolittle, and E. Bapteste. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: A case study of the gamma-proteobacteria. Mol. Biol. Evol. 23:1019–1030.

Vinh le, S., and A. Von Haeseler. 2004. Iqpnni: Moving fast through tree space and stopping in time. Mol. Biol. Evol. 21:1565–1571.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Zweig, M. H., and G. Campbell. 1993. Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. Clin. Chem. 39:561–577.