Syst. Biol. 56(3):389–399, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701397643

Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies

NAIARA RODRÍGUEZ-EZPELETA,[1] HENNER BRINKMANN,[1] BÉATRICE ROURE,[1] NICOLAS LARTILLOT,[2]
B. FRANZ LANG,[1] AND HERVÉ PHILIPPE[1]

[1]*Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada; E-mail: Herve.Philippe@UMontreal.CA (H.P.)*
[2]*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161, rue Ada, 34392 Montpellier Cedex 5, France*

Abstract.—Genome-scale data sets result in an enhanced resolution of the phylogenetic inference by reducing stochastic errors. However, there is also an increase of systematic errors due to model violations, which can lead to erroneous phylogenies. Here, we explore the impact of systematic errors on the resolution of the eukaryotic phylogeny using a data set of 143 nuclear-encoded proteins from 37 species. The initial observation was that, despite the impressive amount of data, some branches had no significant statistical support. To demonstrate that this lack of resolution is due to a mutual annihilation of phylogenetic and nonphylogenetic signals, we created a series of data sets with slightly different taxon sampling. As expected, these data sets yielded strongly supported but mutually exclusive trees, thus confirming the presence of conflicting phylogenetic and nonphylogenetic signals in the original data set. To decide on the correct tree, we applied several methods expected to reduce the impact of some kinds of systematic error. Briefly, we show that (i) removing fast-evolving positions, (ii) recoding amino acids into functional categories, and (iii) using a site-heterogeneous mixture model (CAT) are three effective means of increasing the ratio of phylogenetic to nonphylogenetic signal. Finally, our results allow us to formulate guidelines for detecting and overcoming phylogenetic artefacts in genome-scale phylogenetic analyses. [Compositional heterogeneity; data removal; eukaryotic phylogeny; inconsistency; long-branch attraction; nonphylogenetic signal; phylogenomics; systematic error.]

The use of large multigene data sets to infer phylogenetic trees (phylogenomics) has been successfully applied to resolve evolutionary questions for which single-gene phylogenies failed (Bapteste et al., 2002; Delsuc et al., 2005, 2006; Madsen et al., 2001; Murphy et al., 2001; Philippe et al., 2005a; Qiu et al., 1999; Rodríguez-Ezpeleta et al., 2005; Soltis et al., 1999). This increase in resolution results from the reduction of sampling error through the addition of phylogenetically informative positions. However, higher statistical support does not necessarily lead to more accurate results, because the potential for systematic errors also grows with the increasing size of data sets, which in some cases may lead to strongly supported but incorrect phylogenies (Brinkmann et al., 2005; Jeffroy et al., 2006; Philippe et al., 2004, 2005b; Stefanovic et al., 2004).

In the probabilistic framework (maximum likelihood and Bayesian inference), systematic errors can be traced back to misspecifications in the model of sequence evolution (model violations). Known causes of model violations are across-site rate variation (Yang, 1994), heterotachy (the across-site rate variation through time) (Kolaczkowski and Thornton, 2004; Philippe et al., 2005c; Spencer et al., 2005), site-interdependent evolution (Robinson et al., 2003; Rodrigue et al., 2005), compositional heterogeneity (Foster, 2004; Galtier and Gouy, 1995; Lockhart et al., 1992), and site-heterogeneous nucleotide/amino acid replacement (Lartillot and Philippe, 2004; Pagel and Meade, 2004). In the following, we will call the apparent signal arising from such model violations "nonphylogenetic" signal, as opposed to genuine phylogenetic signal that corresponds to bona fide shared-derived characters.

The impact of model violations on phylogenetic accuracy is greatly exaggerated when multiple substitutions occur at given sites (mutational saturation). In the absence of model violation, mutational saturation would result in random sequences simply leading to poorly resolved trees (but see Susko et al., 2005). In contrast, when the model is violated, systematic error becomes manifest. Because long branches (either due to fast evolutionary rate or long time span) accumulate more multiple substitutions, they are most affected by long branch attraction (LBA), the well-known case of systematic error that provokes the clustering of fast-evolving species regardless of their true phylogenetic relationship (Felsenstein, 1978). Several complementary approaches have been applied to overcome systematic errors such as LBA: (i) increased taxon sampling and improved models of sequence evolution, allowing a more efficient detection of multiple substitutions; and (ii) removal of fast-evolving species (Aguinaldo et al., 1997), genes (Brinkmann et al., 2005; Philippe et al., 2005b), or sequence positions (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Hirt et al., 1999; Ruiz-Trillo et al., 1999).

In this paper, we study the relative contribution of phylogenetic and nonphylogenetic signal to genome-scale phylogenies and explore different methods to overcome systematic error. We use the global eukaryotic phylogeny as a case study for two reasons. First, the eukaryotic diversification is difficult to resolve, possibly because of closely spaced speciation events (Knoll, 1992; Philippe and Adoutte, 1998), implying that the phylogenetic signal would be limited, and second, multiple substitutions are expected given the long time span of eukaryotic evolution, most likely making nonphylogenetic signal significant.

Using a data set of 143 nuclear encoded protein sequences from 37 eukaryotic species, we show that slight deviations in the evolutionary rate or amino acid

composition of the sequences can lead to strongly supported but incorrect phylogenies. This occurs when the phylogenetic signal for a given branch is significantly weaker than the nonphylogenetic signal. Alternatively, when both signals are of equivalent strength, they may counterbalance each other, leading to unresolved trees, even with large data sets. We demonstrate that (i) variations in taxon sampling, (ii) removal of fast-evolving sites, (iii) use of a site-heterogeneous mixture model (Lartillot and Philippe, 2004), and (iv) amino acid coding into functional categories have the potential to overcome some types of systematic errors in genome-scale data sets.

Materials and Methods

Phylogenetic Analyses

The analyses were performed on a previously described data set of 143 nuclear-encoded proteins (30,244 amino acid positions) from 39 eukaryotic species (Rodríguez-Ezpeleta et al., 2005), excluding the two fastest-evolving lineages (*Trichomonas vaginalis* and *Giardia lamblia*). Trees were inferred using maximum parsimony (MP), Bayesian inference (BI), and maximum likelihood (ML) methods. The alignments (including corresponding trees) have been submitted to TreeBASE under accession numbers SN3166-13372 to SN3166-13377.

Heuristic Analyses

MP analyses were performed using PAUP* (Swofford, 2000), with tree bisection and reconnection search and 10 random additions of species. The support was evaluated based on 1,000 bootstrap replicates. BI analyses were conducted using MrBayes 3.0 b4 (Ronquist and Huelsenbeck, 2003) or PhyloBayes (http://www.lirmm.fr/mab/article.php3?id_article=329). MrBayes analyses were performed with the WAG amino acid replacement matrix (Whelan and Goldman, 2001), gamma-distributed rates across sites (four discrete categories), and stationary amino acid frequencies estimated from the data set (WAG+F+Γ4 model). Three independent analyses with 120,000 generations gave identical results. PhyloBayes analyses were performed with the CAT mixture model, which accounts for across-site heterogeneities in the amino acid replacement process (Lartillot and Philippe, 2004). Two independent runs were performed with a total length of 2500 cycles (250 topological moves per cycle), with the same operators as in Lartillot et al. (2006). The first 500 points were discarded as burn-in, and the posterior consensus was computed on the 2000 remaining trees. Preliminary ML analyses were performed on the concatenated data set using heuristic searches with PhyML 2.4 (Guindon and Gascuel, 2003) and TreeFinder (Jobb et al., 2004) with the WAG+F+Γ4 model. The support was evaluated based on 100 bootstrap replicates.

Exhaustive Analyses

The probability of getting trapped in a local minimum during heuristic topology searches is high for large data sets (Salter, 2001), but an exhaustive search is impossible in our case given the large number of possible topologies for 37 species (10^{49}). This problem was addressed by constraining relationships supported by consistently more than 95% bootstrap values (MP and ML) and 1.0 posterior probability (BI) (opisthokonts—animals, choanoflagellates and fungi, red algae, green plants, glaucophytes, apicomplexans, stramenopiles and kinetoplastids). This reduces the number of topologies to be exhaustively analyzed to 135,135. To further alleviate computational cost and memory usage, we proceeded in two steps. First, exhaustive ML analyses without taking rate across-site variation into account were performed with PROTML (Adachi and Hasegawa, 1996) and the JTT amino acid replacement matrix (Jones et al., 1992) for each protein separately (for details, see Rodriguez-Ezpeleta et al., 2005). The resulting 135,135 tree topologies were sorted by likelihood value, and the top 1733 trees were selected. These trees were augmented by sampling every 500th subsequent topology, for a total of 2000 trees. For these 2000 trees, likelihood values were calculated with TREE-PUZZLE (Schmidt et al., 2002) and the concatenated WAG+F+Γ4 model (all parameters estimated for the concatenated data set). We verified that retention of the 1733 top ranking topologies was sufficient: first, the correlation between the likelihood values obtained with the separate JTT+F and the concatenated WAG+F+Γ4 models for the 2000 selected topologies is excellent (R^2= 0.9693; Fig. S1 [supplementary figures available online at www.systematicbiology.org]); second, the order of topologies is almost identical with and without considering rates across sites; and third, the nine best topologies from the separate JTT+F analysis receive a total of 98% of the RELL bootstrap support (Kishino et al., 1990) (the 83 best topologies receive 100% of the RELL bootstrap support). Indeed, retaining 100 topologies gives virtually identical results (not shown). In order to estimate statistical support for each branch, the RELL bootstrap method (Kishino et al., 1990) was used. In brief, site-wise likelihood values were calculated with PAML (Yang, 1997) with the concatenated WAG+F+Γ4 model and used to perform RELL bootstrap analyses with 10,000 replicates.

The relationship between the number of sequence positions and the bootstrap support values (BVs) was calculated as described (Lecointre et al., 1994). Briefly, different numbers of positions (3000, 6000, etc.) were randomly drawn from the complete data set 100 times. RELL bootstrap values (100 replicates) were then computed for each of the 100 samples and for each size fraction (site-wise likelihoods were not recomputed for each sample for obvious computation time reasons but are expected to be similar with this large number of positions; see below). The average of the BV of all branches for each size fraction was plotted against its size.

Removal of Fast-Evolving Sites

Fast-evolving sites were identified using a modification of the method proposed by Ruiz-Trillo et al. (1999)

and Burleigh and Mathews (2004). Instead of eliminating sites according to the discrete gamma category to which they most likely belong, they were eliminated according to their site-wise rates calculated by PAML (i.e., weighted-average rates over all categories with the weights given by the posterior probabilities of each category) on the concatenated data set for each topology. Sites were then sorted according to (i) the rates estimated on a given topology or to (ii) the mean of the rates estimated on all topologies. Then, fast-evolving sites were progressively removed in steps of 1000. RELL bootstrap analyses (1000 replicates) were performed after each step, and the resulting values plotted against the alignment size.

The computational burden associated to site removal is only circumvented if the BVs are computed using the RELL method. However, two important assumptions of this method may be violated if too many sites are removed: (i) the parameters of the model estimated on the complete data set (in particular, branch lengths) should remain similar for the reduced data set and (ii) the topological constraints imposed should remain valid. First, the constraints were verified after the removal of 15,000 and 20,000 sites by performing heuristic analyses with TreeFinder; and second, the parameters and the site-wise likelihoods were reestimated on these two data sets. After the removal of 15,000 sites (half of the data set) all constraints are still respected, and the results obtained with and without reestimating site-wise likelihood values are similar (the correlation coefficient between BVs is 0.86). However, after removal of 20,000 sites, some of the constraints are no longer supported (e.g., the sister group of apicomplexans and ciliates), and the RELL bootstrap values obtained before and after parameter reestimation differ substantially. We thus stopped after the removal of 15,000 sites in all analyses.

Testing for Saturation

The saturation of the alignments was measured by plotting the number of observed differences (p distances) against the number of substitutions that are computed as patristic distances (in our case, derived from the ML tree) using TREEPLOT of the MUST package (Philippe, 1993; Philippe et al., 1994). Both distance matrices were compared, and the slope of the graph was calculated using the COMP_MAT program in the MUST package. The greater the number of inferred substitutions with respect to the number of observed differences (small slope), the greater the saturation of the data (Jeffroy et al., 2006).

Compositional Heterogeneity

The amino acid composition bias of the species in the data set was visualized by assembling a 37×20 matrix containing the percentage of each amino acid per species using the NET program from the MUST package (Philippe, 1993). This matrix is displayed as a two-dimensional plot in a principal component analysis (PCA), as implemented in the SAS program (SAS, 1999). To calculate the overall compositional bias in the data, the Bowker's test for compositional symmetry (Ababneh et al., 2006; Bowker, 1948) was applied. Bowker's values were calculated for each pair of sequences and the median value was computed. Large Bowker's values indicate strong heterogeneity in the data set, whereas lower Bowker's values indicate that the sequence composition is homogeneous (note that the phylogenetic dependency among all Bowker's values is not corrected for here).

Two attempts to reduce the potential impact of compositional bias were performed, by (i) constructing neighbour-joining trees based on LogDet+Γ pairwise distances, calculated with the LDDist perl module (Tholleson, 2004) and using the rate categories estimated by TREE-PUZZLE; and (ii) recoding the data using the common six groups of amino acids that usually replace one another (Hrdy et al., 2004). To allow for a general-time-reversible (GTR) matrix implemented in most programs, the data set was recoded to four categories instead of six, by combining aromatic (FYW) and hydrophobic (MVIL) amino acids and coding the rare cysteine as missing data. The four amino acid categories were named A, T, G, and C, respectively, and the parameters of the GTR matrix were estimated by PAUP. The 2000 best topologies from the exhaustive search were analyzed by TREE-PUZZLE with a GTR+F+Γ4 model. RELL bootstrap (10,000 replicates) analyses were performed as described above. The constraints were verified after the recoding with heuristic ML analyses using TreeFinder.

Results and Discussion

Phylogenomic Analyses Do Not Resolve Every Branch

Figure 1 shows the ML tree based on 143 nuclear protein-coding genes (30,244 amino acid positions) from 37 eukaryotic species. The monophyly of all major eukaryotic groups and the relationships within them are recovered with 100% bootstrap support value (BV) and are in agreement with current knowledge of eukaryotic evolution (Baldauf et al., 2000; Simpson and Roger, 2004), underlining that the use of a large number of genes notably improves overall statistical support. Only four branches receive BVs below 100%. Among them, the monophyly of primary photosynthetic eukaryotes or Plantae (green plants, rhodophytes, and glaucophytes) requires special attention. This grouping has already been suggested based on genomic features and molecular phylogenies of plastid and nuclear proteins (Cai et al., 2003; Huang and Gogarten, 2006; McFadden and van Dooren, 2004; Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005); however, with a particular taxon sampling (Fig. 1), it only receives statistically nonsignificant support (64% BV).

Unsupported trees are usually attributed to a lack of phylogenetic information in the data, suggesting that the addition of more genes or positions will increase resolution (Bapteste et al., 2002; Rodríguez-Ezpeleta et al., 2005; Rokas et al., 2003; Saitou and Nei, 1986). Therefore, we studied the variation of the BVs obtained for the monophyly of Plantae with respect to the number of amino acid positions considered. As shown in Figure 2 (open triangles), the BVs rapidly increase as more positions are added. But when more than about 10,000

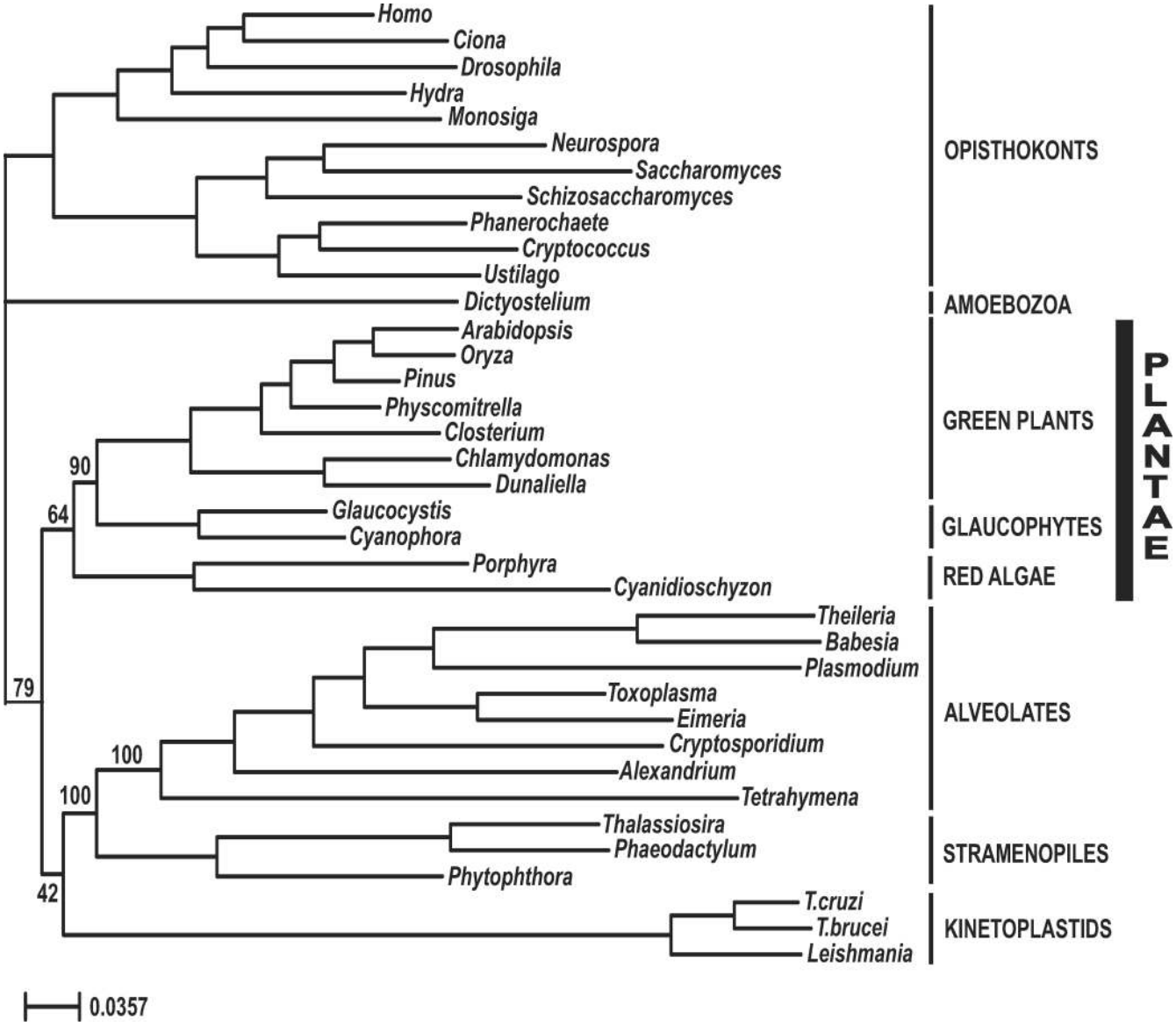

FIGURE 1. Eukaryotic phylogeny based on 143 nuclear-encoded proteins (30,244 amino acid positions) inferred by exhaustive ML analysis with the concatenated WAG+F+Γ4 model. The same topology is obtained with PhyML, TreeFinder, and MrBayes. Numbers indicate bootstrap values obtained by analyzing 10,000 RELL replicates on the exhaustive ML analysis. Branches without values are supported by BVs of 100 and posterior probabilities of 1.0 in the ML (PhyML and TreeFinder) and BI (MrBayes) analyses, respectively, and were constrained in the exhaustive analysis. The scale bar denotes the estimated number of amino acid substitutions per site.

amino acid positions are considered, the BVs attain a plateau, suggesting that the addition of more data (even of complete genome sequences) will most likely not lead to a statistically significant support for the monophyly of Plantae, given this taxon sampling and this tree reconstruction method. In fact, an alternative grouping, the sister group relationship of red algae and kinetoplastids (Fig. 2; closed circles), displays very similar behavior, rising rapidly to a plateau of 40% BV.

The shape of the curves obtained in Figure 2 suggests that the unsupported monophyly of Plantae is not due to a lack of phylogenetic signal. Rather, it seems as if two competing signals exist in the data: one that supports the monophyly of Plantae and another one that supports a sister-group relationship between red algae and kinetoplastids.

Coexistence of Phylogenetic and Nonphylogenetic Signal in the Data

Because kinetoplastids present the longest unbroken branch in the data set (Fig. 1), the hypothesis of an LBA artefact as the cause for their clustering with red algae can be advanced. To test if the two red algae (*Cyanidioschyzon* and *Porphyra*; both have moderate evolutionary rate differences) are differently affected by this artefact, two data sets were created, using either *Porphyra* or *Cyanidioschyzon* as the single representative of the red algae.

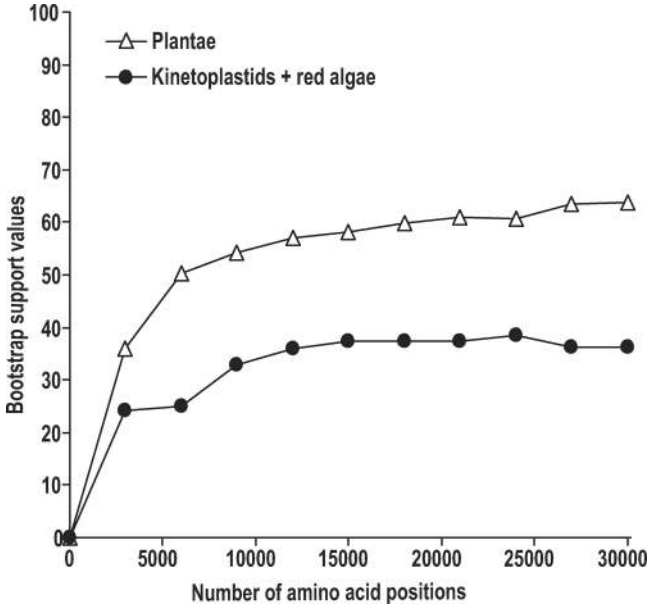

FIGURE 2. Bootstrap values for the monophyly of Plantae (open triangles) and the sisterhood of red algae and kinetoplastids (closed circles) as a function of the number of amino acid positions. Bootstrap values were obtained by sampling different numbers of positions (3000, 6000, etc.) 100 times and by averaging the RELL bootstrap values (100 replicates) for samples of the same size.

Surprisingly, the use of one or the other red algae has drastic effects on the outcome. With *Porphyra* as the sole red algal representative, the BV for Plantae raises from 64% to 99% (Fig. 3a), whereas with *Cyanidioschyzon* alone, the support for Plantae drops to 0% and the support for the sisterhood of red algae and kinetoplastids raises to 100% (Fig. 3b). Because we share the view with others that red algae are indisputably monophyletic (e.g., Ragan and Gutell, 1995; see also Fig. 1), one of the two trees in Figure 3 has to be wrong. Because *Cyanidioschyzon* evolves somewhat faster than *Porphyra*, our working hypothesis is that the monophyly of Plantae observed in Figure 3a is the product of genuine phylogenetic signal, whereas the grouping of red algae and kinetoplastids (Fig. 3b) is an LBA artefact.

As *Cyanidioschyzon* evolves only 1.25 times faster than *Porphyra* (Fig. 1), the radical difference in the resulting tree topologies (Fig. 3a, b) may seem surprising. We posit that this can be explained by the large number of amino acid positions in this data set. More than 15,000 amino acid positions (Fig. 3d) are required to recover the sister-group of *Cyanidioschyzon* and kinetoplastids with BV >95%, suggesting that the nonphylogenetic signal is weak; however, the phylogenetic signal for the monophyly of Plantae is as weak (Fig. 3c).

Testing the LBA Hypothesis Using Differences in Taxon Sampling

If the grouping of *Cyanidioschyzon* and the kinetoplastids is due to LBA, this artefact should be reproduced with other long unbroken branches in this data set. To explore this hypothesis, three combinations of taxa were created that induce long unbroken branches. Starting from the data set of Figure 3b, the kinetoplastids were removed and (i) *Saccharomyces* was kept as the only representative of the *Dictyostelium*/opisthokont clade, and either (ii) *Theileria* and *Phytophthora* or (iii) *Plasmodium* and *Phytophthora* were kept as the only representatives of alveolates and stramenopiles, respectively.

In all cases, only *Cyanidioschyzon* is attracted to the longest unbroken branch (Fig. 4). Importantly, Plantae remain monophyletic in these three cases when *Porphyra* is used (Fig. S2). This confirms that the grouping of kinetoplastids and *Cyanidioschyzon* is due to LBA. Surprisingly, the grouping of *Plasmodium* and *Cyanidioschyzon* receives only 66% BV (Fig. 4c), whereas the grouping of *Theileria* and *Cyanidioschyzon* (Fig. 4b) has 90% BV. Interestingly, in a principal component analysis (Fig. 5), the amino acid composition of *Cyanidioschyzon* is most similar to that of *Saccharomyces* and kinetoplastids, less to *Theileria*, and least to *Plasmodium*—the species with the most extreme genomic A+T content (80.6%). Therefore, even if the two alveolates (*Theileria* and *Plasmodium*) have almost the same evolutionary rate (Fig. 1), the extreme compositional bias in *Plasmodium* appears to have an additional effect on the bootstrap support (Fig. 4b, c).

Extracting Phylogenetic Signal by Removing Fast-Evolving Sites

Because fast-evolving sites are more likely to be saturated and prone to accumulation of nonphylogenetic signal, a progressive removal of such sites should decrease artefacts caused by model violations (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Olsen, 1987; Ruiz-Trillo et al., 1999). We studied the impact of the fast sites in our data set by progressively removing blocks of the fastest-evolving sites.

The estimation of site-specific rates requires the knowledge of a tree topology. To avoid circularity, we used the best (ML) topology obtained with a data set that does not include the red algae (which we cannot place with confidence with all the data). The experiment was performed on the data sets from Figures 3b, 4a to c. In three cases, the removal of the fast evolving sites strengthens the support for the monophyly of Plantae and lowers the one for the alternative position (Fig. 6a, c, d), confirming that the removal of the fast-evolving sites increases the ratio of phylogenetic to nonphylogenetic signal (Brinkmann and Philippe, 1999; Brochier and Philippe, 2002). With *Saccharomyces* as the only representative of opisthokonts and Amoebozoa, the removal of the fastest-evolving sites is insufficient to recover the monophyly of Plantae, although a small increase in the BV is observed (Fig. 6b). The number of sites that need to be removed to recover this relationship is different in each case, which may result from different levels of nonphylogenetic signal in various data sets.

For the experiments described above, a tree topology without red algae was used to calculate site-wise rates (to avoid introduction of bias). The procedure is justified in this special case, where a single taxon is added to

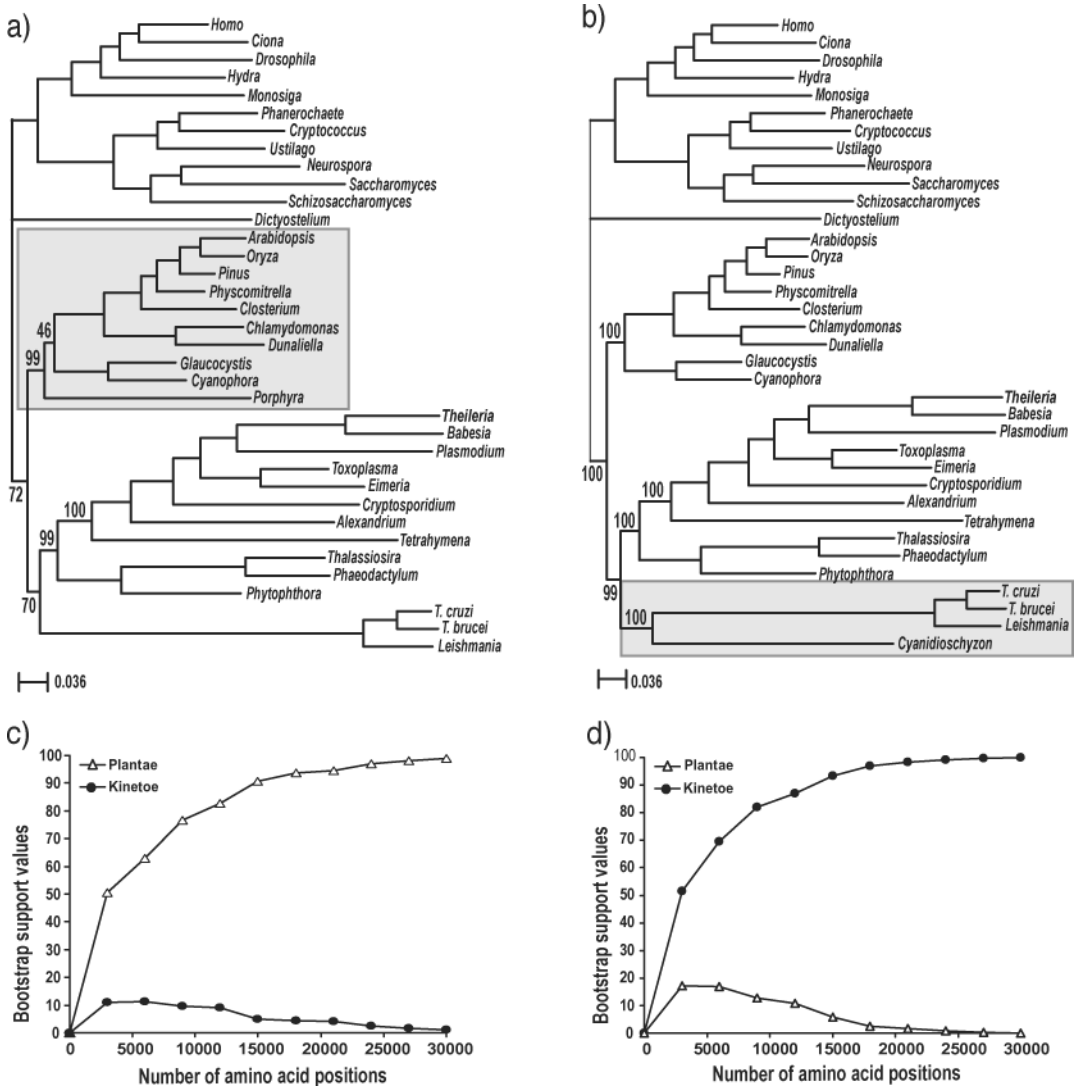

FIGURE 3. Alternative topologies obtained as described in Figure 1 when only *Porphyra* (a) or only *Cyanidioschyzon* (b) was used to represent the red algae. No value above branch indicates that the corresponding node was supported at 100% BV in the ML analyses with PhyML and TreeFinder and was constrained in the exhaustive analysis. Grey shaded areas indicate the alternative positions of red algae. For each data set, the bootstrap values of the two alternative positions for red algae were plotted against the number of amino acid positions (c and d).

an otherwise unquestioned topology but should probably not be applied when more complex changes are expected. To test if the choice of tree topology significantly affects the estimation of site-wise rates, results were compared for the red algae + kinetoplastids (Fig. 3b) and the Plantae topology (Fig. 3a). When the rates were estimated on the red algae + kinetoplastids topology, the removal of the fastest-evolving sites does not improve phylogenetic accuracy (Fig. S3); in contrast, if the rates were estimated on the Plantae topology, the removal of even fewer sites than in Figure 6 leads to recovery of the correct topology (Fig. S4). Evidently, the specific topology used to estimate the rates heavily influences the results. As a solution to this problem, we propose to use the mean site-wise rates estimated for a given set of best topologies. In our specific example, with the 2000 topologies, results are virtually identical to the experiment in which a tree without red algae was used (Fig. S5). This "mean rate approach" is an interesting avenue that deserves further investigation.

Fast-Evolving Sites Are Mutationally Saturated and Compositionally Biased

For each of the nonoverlapping windows of 1000 sites that have been progressively removed, the mutational saturation and the compositional bias were studied. As expected, the mutational saturation (grey line in Fig. 7) is tightly correlated to the evolutionary rates, confirming that the fast-evolving sites are the most saturated. Because the effects of model violations are more evident in mutationally saturated sites, the removal of the

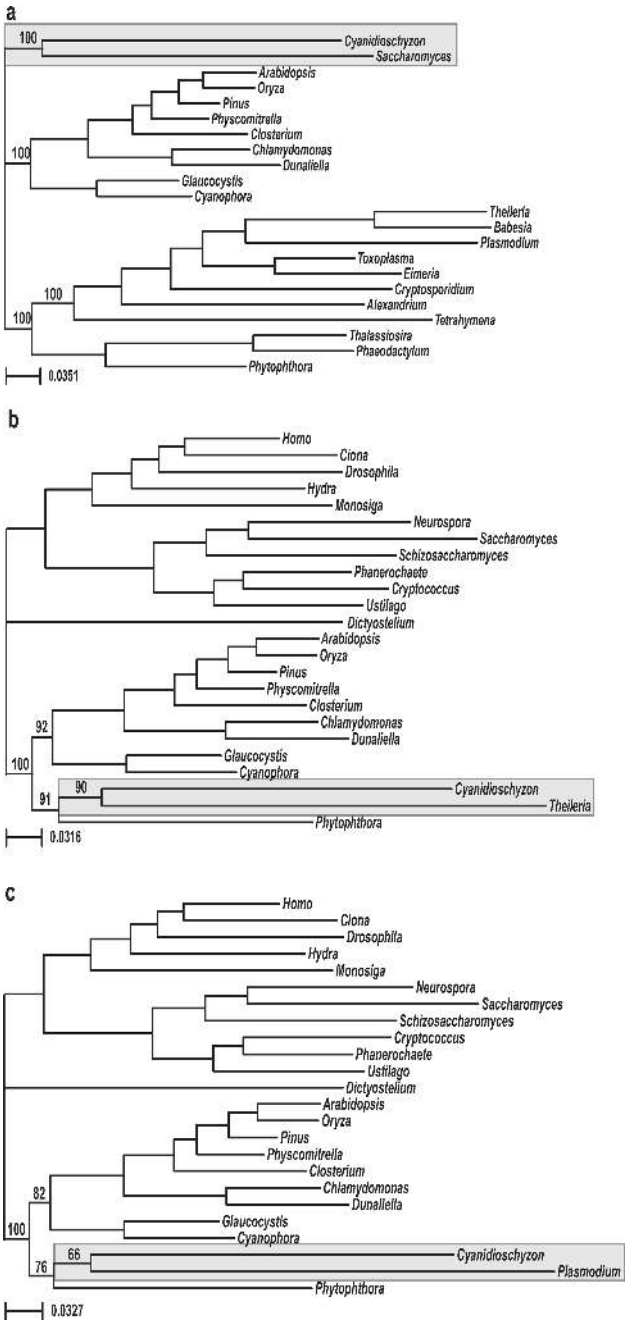

FIGURE 4. Same analyses as in Figure 1 but with selected combinations of taxon samplings that are likely to induce an LBA artefact. In all cases, only *Cyanidioschyzon* was used to represent the red algae, and the kinetoplastids were excluded from the data set; (a) using *Saccharomyces* as the only representative of the opisthokonts and Amoebozoa; using either (b) *Theileria* and *Phytophthora* or (c) *Plasmodium* and *Phytophthora* as the representatives of alveolates and stramenopiles, respectively. No value above branches indicates that the corresponding node was supported at 100% BV in the ML analyses with PhyML and TreeFinder and was constrained in the exhaustive analysis. Grey shaded areas indicate the position *Cyanidioschyzon*.

fastest-evolving sites efficiently overcomes systematic errors. We also measured a well-known source of model violation, the compositional heterogeneity among lineages. For each of the successively removed blocks of

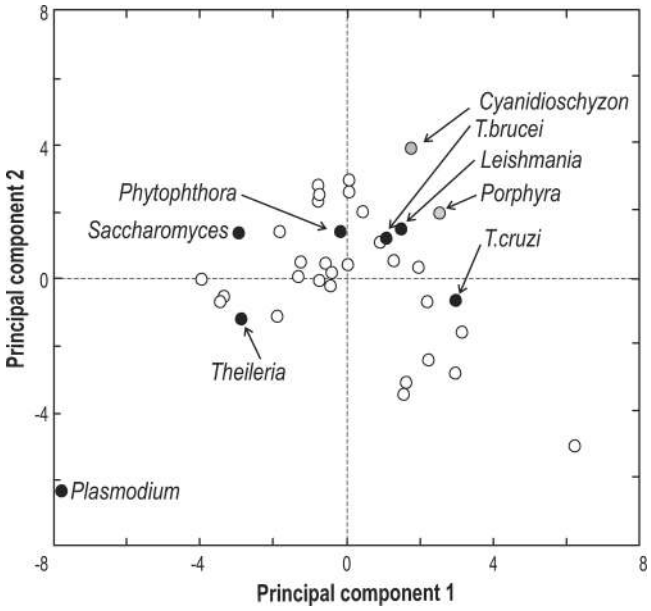

FIGURE 5. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The variances that explain the two first axes are respectively 32% and 25%. Grey circles denote the two red algae and black circles are other relevant species used in previous analyses.

1000 positions, the Bowker's test for compositional symmetry was computed (black line in Fig. 7). Interestingly, the compositional heterogeneity is tightly correlated with the rate of the sites: the most saturated sites are the most compositionally biased. Therefore, by removing the fast-evolving sites, we most likely overcome systematic error due to compositional heterogeneity. Accordingly, other sources of model violations may also be decreased by fast-evolving site removal, and this question deserves further studies.

The Effects of Model Violations

Another kind of model violation that may result in phylogenetic artefacts is the heterogeneity of the amino acid replacement process across sites (Baurain et al., 2006; Koshi and Goldstein, 2001; Lartillot et al., 2006; Lartillot and Philippe, 2004; Pagel and Meade, 2004). Most sites of a protein show substitutions among a small set of two to four biochemically equivalent amino acids (Miyamoto and Fitch, 1996). However, homogeneous models inherently assume that, under maximal saturation, all 20 amino acids are likely to be observed at any given site with probabilities equal to the equilibrium frequencies. As a result, the probability of convergence is strongly underestimated by standard models of evolution (Chang, 1996; Felsenstein, 2004).

The site-heterogeneous mixture model, CAT (Lartillot and Philippe, 2004), was applied to the various taxon samplings previously studied. The monophyly of Plantae was recovered in all but two cases even when *Cyanidioschyzon* is the only red alga. In particular, in two cases

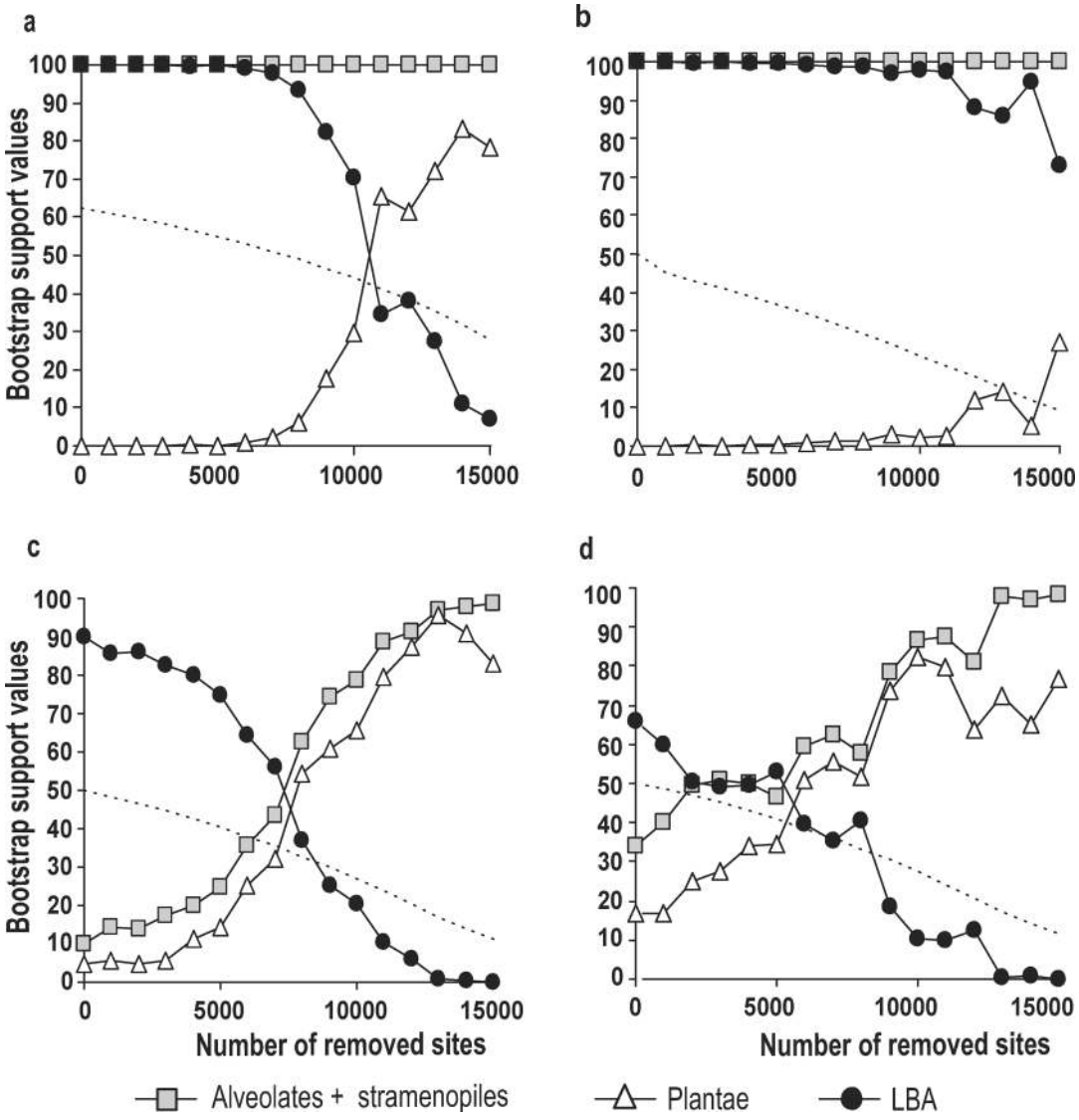

FIGURE 6. Progressive removal of fast-evolving sites from the data sets of Figures 3b (graph a), 4a (graph b), 4b (graph c), and 4c (graph d). The site-specific rates were calculated with the best ML topology from which *Cyanidioschyzon* was excluded. Dotted line represents the number of parsimony-informative positions.

where the homogeneous model fails (Fig. 4b, c), the inference with a more complex model is not sensitive to systematic error: when *Theileria* or *Plasmodium* are the only representatives of alveolates, Plantae where supported by a posterior probability (pp) of 0.98 and 0.85, respectively. Nevertheless, the monophyly of Plantae was not recovered (pp = 0) in the presence of kinetoplastids or of *Saccharomyces* as LBA attractors. Therefore, site-specific substitution pattern is not the only cause of the observed artefacts.

An alternative potential source of model violation is the nonstationarity of the amino acid replacement process, known to affect our data set (Fig. 5). Under a stationary model, where the same amino acid or nucleotide composition is assumed along the tree, compositional heterogeneity may drastically mislead phylogenetic reconstruction (Hasegawa and Hashimoto, 1993; Hendy and Penny, 1989; Lockhart et al., 1992; Phillips et al., 2004). Although models have been developed to overcome this violation (e.g., Foster, 2004; Galtier and Gouy, 1995; Yang and Roberts, 1995; Blanquart and Lartillot, 2006), they are computationally demanding and have implementation limitations and are therefore of limited value. Other ways to overcome nonphylogenetic signal due to compositional heterogeneity have been reported, such as the LogDet method (Lake, 1994; Lockhart et al., 1994) and the RY (Woese et al., 1991) or Dayhoff (Hrdy et al., 2004) coding for nucleotides and amino acids.

Interestingly, amino acid coding into functional categories has an impact on both kinds of model violations mentioned above; i.e., compositional effects and the

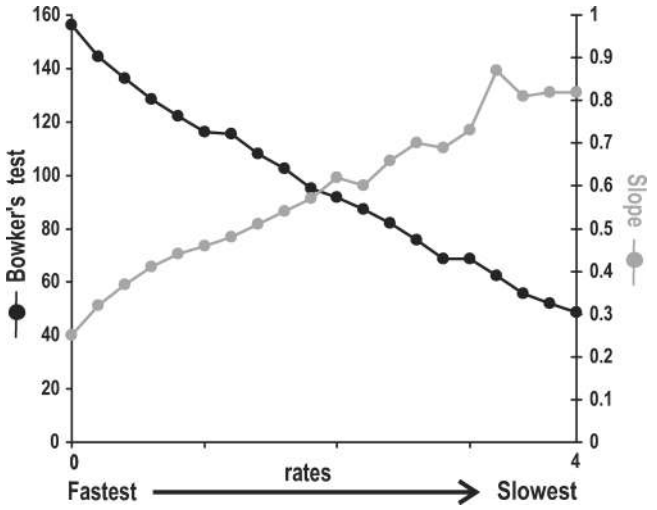

FIGURE 7. Amino acid compositional bias and level of saturation observed in blocks of 1000 positions when progressively removing sites from fast to slow. For each block, the Browker's test for amino acid composition (black) and the correlation between the observed differences and estimated substitutions (grey) were performed.

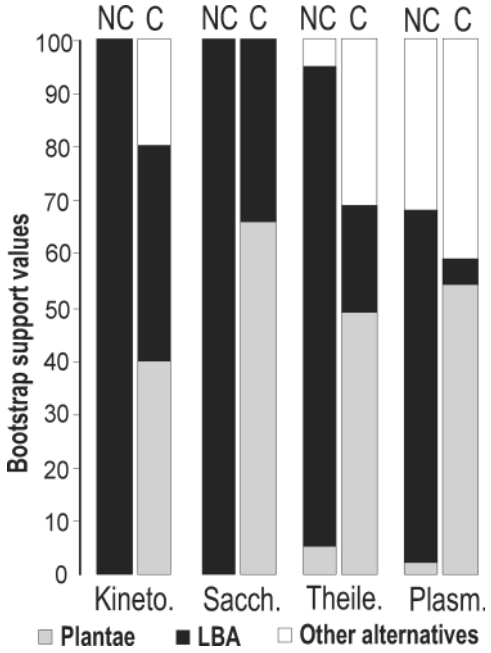

FIGURE 8. Differences in bootstrap support without (WAG+F+Γ4 model) and with amino acid recoding (GTR+F+Γ4 model). Four data sets including *Cyanidioschyzon* as the only red algae were analyzed before (NC) and after the coding (C). Support for the monophyly of Plantae, grey; misplacement of the red algae as shown in Figures 3b (Kineto.), 4a (Sacch.), 4b (Theile.), and 4c (Plasm.), black.

expected number of amino acids per position. First, it alleviates compositional bias (Phillips et al., 2004; Woese et al., 1991). For example, lysine (K) and arginine (R) are two easily interchangeable amino acids whose codons differ at a single position (AAR and AGR, respectively) and that are preferred in AT- and GC-rich genomes, respectively. Hence, coding pairs or groups of amino acids such as K and R as a single character state should compensate for these biases. Second, the recoding will also alleviate the problem of homoplasies that occur in peaked biochemical profiles by reducing the number of character states from 20 to 4.

Applied to our data set, the LogDet method failed to recover the expected tree topology, and a strong LBA artefact unites alveolates and kinetoplastids to the exclusion of stramenopiles, a grouping that attracts *Cyanidioschyzon*. In fact, it has already been suggested that the LogDet method may fail in the presence of rate heterogeneity among sites or lineages (Conant and Lewis, 2001). Instead, a modification of the Dayhoff coding (Hrdy et al., 2004; see Material and Methods) increases the support for Plantae while decreasing the attraction of *Cyanidioschyzon* with long unbroken branches (Fig. 8) in all four cases (Figs. 3b, 4a to c). Importantly, with amino acid recoding, Plantae monophyly was recovered with *Saccharomyces* as the only representative of Opisthokonts, when all other methods failed.

Altogether, the overall pattern suggests that the artefacts observed in this data set are mainly caused by a combination of compositional bias and site-heterogeneity that operate at different levels, depending on the attractor: a site-heterogeneity violation, dominant in the case of *Plasmodium* and *Theileria*, and possibly compositional bias with kinetoplastids and *Saccharomyces*. As discussed above, recoding is efficient in alleviating both sources of systematic error simultaneously, although it reduces the phylogenetic signal considerably.

CONCLUSION

The common view that using genome-scale data sets is a universal remedy for resolving phylogenetic questions (e.g., Rokas et al., 2003) is inaccurate. Tree reconstruction artefacts that are invisible in single-gene phylogenies may become dominant in large data sets (Jeffroy et al., 2006). Depending on the relative contribution of phylogenetic and nonphylogenetic signal, certain genome-scale data sets may either lead to predicting incorrect tree topologies with confidence, or one or more branches remain unresolved whatever the data size.

The identification of "misbehaving" species that contribute an unproportional fraction of nonphylogenetic signal is possible through variations in taxon sampling. Removal of these species from the data set has been common practice to overcome some phylogenetic artefacts. Alternatively, more general approaches include data recoding, removal of fast-evolving sites, or the use of more realistic models of sequence evolution. Yet, current implementations of these procedures will either eliminate much of the phylogenetic signal, or are impracticable in terms of computational load. In practical terms, we therefore recommend a combined application of all methods that will overcome at least some of the well-known types of systematic errors. Evidently, these approaches cannot address all kinds of systematic error present in a data set;

for example, none of the techniques applied here detect or overcome heterotachy (rate heterogeneity across sites through time).

Ultimately, the development of more sophisticated models of sequence evolution that address simultaneously the different kinds of systematic biases will reduce the requirement for intense user intervention by making best use of phylogenetic signal.

ACKNOWLEDGMENTS

We wish to thank Denis Baurain for assistance with the tests for amino acid composition and helpful comments on a previous version of the manuscript, and Pablo Vinuesa, Frank (Andy) Anderson, Andrew J. Roger, and two anonymous reviewers for useful suggestions. This work has been supported by operating and equipment funds from Genome Quebec/Canada. B.F.L. and H.P. acknowledge the program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR) for salary and interaction support and the Canada Research Chairs Program and the Canadian Foundation for Innovation (CFI) for salary and equipment support. N.R.E. has been supported by Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación (Government of Basque Country) and B.R. by Bourses d'Excellence biT (CIHR).

REFERENCES

Ababneh, F., L. S. Jermiin, C. Ma, and J. Robinson. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231.

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28:1–150.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493.

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 H. Philippe. 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.

Baurain, D., H. Brinkmann, and H. Philippe. 2006. Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? Mol. Biol. Evol. 24:6–9.

Blanquart, S., and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23:2058–2071.

Bowker, A. H. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43:572–574.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

Brochier, C., and H. Philippe. 2002. Phylogeny: A nonhyperthermophilic ancestor for Bacteria. Nature 417:244.

Burleigh, J. G., and S. Mathews. 2004. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Cai, X., A. L. Fuller, L. R. McDougald, and G. Zhu. 2003. Apicoplast genome of the coccidian *Eimeria tenella*. Gene 321:39–46.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134:189–215.

Conant, G. C., and P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol. Biol. Evol. 18:1024–1033.

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the Tree of Life. Nat. Rev. Genet. 6:361–375.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA 92:11317–11321.

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., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. USA 96:580–585.

Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy, and T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618–622.

Huang, J., and J. P. Gogarten. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet. 22:361–366.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: The beginning of incongruence? Trends Genet. 22:225–231.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: A powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4:18.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

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.

Knoll, A. H. 1992. The early evolution of eukaryotes: A geological perspective. Science 256:622–627.

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

Koshi, J. M., and R. A. Goldstein. 2001. Analyzing site heterogeneity during protein evolution. Pac. Symp. Biocomput. 6:191–202.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455–1459.

Lartillot, N., H. Brinkmann, and H. Philippe. 2006. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(Suppl 1):S4.

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.

Lecointre, G., H. Philippe, H. L. Van Le, and H. Le Guyader. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol. Phylogenet. Evol. 3:292–309.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34:153–162.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

McFadden, G. I., and G. G. van Dooren. 2004. Evolution: Red algal genome affirms a common origin of all plastids. Curr. Biol. 14:R514–R516.

Miyamoto, M. M., and W. M. Fitch. 1996. Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst. Biol. 45:568–575.

Moreira, D., H. Le Guyader, and H. Philippe. 2000. The origin of red algae and the evolution of chloroplasts. Nature 405:69–72.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, M. S. Springer, D. J. Kao, R. W. DeBry, R. Adkins, and H. M. Amrine. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

Olsen, G. J. 1987. Earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp. Quant. Biol. 52:825–837.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Philippe, H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. 21:5264–5272.

Philippe, H., and A. Adoutte. 1998. The molecular phylogeny of Eukaryota: Solid facts and uncertainties. Pages 25–56 *in* Evolutionary relationships among Protozoa (G. Coombs, K. Vickerman, M. Sleigh, and A. Warren, eds.). Kluwer, Dordrecht.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol. Biol. Evol. 22:1246–1253.

Philippe, H., U. Sörhannus, A. Baroin, R. Perasso, F. Gasse, and A. Adoutte. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. 7:247–265.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.

Ragan, M., and R. Gutell. 1995. Are red algae plants? Bot. J. Linn. Soc. 118:81–105.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20:1692–1704.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217.

Rodríguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe, and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Curr. Biol. 15:1325–1330.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Ruiz-Trillo, I., M. Riutort, D. T. Littlewood, E. A. Herniou, and J. Baguna. 1999. Acoel flatworms: Earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919–1923.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. 24:189–204.

Salter, L. A. 2001. Complexity of the likelihood surface for a large DNA data set. Syst. Biol. 50:970–978.

SAS. 1999. SAS/STAT user's guide. Version 8.12. SAS Institute, Cary, NC.

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.

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

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.

Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol. Biol. 4:35.

Susko, E., M. Spencer, and A. J. Roger. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. J. Mol. Evol. 61:351–359.

Swofford, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony and other methods. Version 4b10. Sinauer Associates, Sunderland, Massachusetts.

Thollesson, M. 2004. LDDist: A Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics 20:416–418.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco. 1991. Archaeal phylogeny: Reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14:364–371.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451–458.

First submitted 21 July 2006; reviews returned 17 October 2006; final acceptance 28 November 2006
Associate Editor: Frank Anderson