Gene Sampling Can Bias Multi-Gene Phylogenetic Inferences: The Relationship between Red Algae and Green Plants as a Case Study

Yuji Inagaki,*† Yoshihiro Nakajima,*† Mitsuhisa Sato,*† Miako Sakaguchi,*† and Tetsuo Hashimoto*†

*Center for Computational Sciences, Tsukuba, Ibaraki, Japan; †Institute for Biological Sciences, Tsukuba, Ibaraki, Japan; and ‡Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki, Japan

The monophyly of Plantae including glaucophytes, red algae, and green plants (green algae plus land plants) has been recovered in recent phylogenetic analyses of large multi-gene data sets (e.g., those including >30,000 amino acid [aa] positions). On the other hand, Plantae monophyly has not been stably reconstructed in inferences from multi-gene data sets with fewer than 10,000 aa positions. An analysis of 5,216 aa positions in Nozaki et al. (Nozaki H, Iseki M, Hasegawa M, Misawa K, Nakata T, Sasaki N, Watanabe M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol Biol Evol. 24:1592–1595.) strongly rejected the monophyly of Plantae, whereas Hackett et al. (Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol Biol Evol. 24:1702–1713.) robustly recovered the Plantae clade in an analysis of 6,735 aa positions. We suspected that the significant incongruity observed between the two studies was attributable to a bias generally overlooked in multi-gene phylogenetic estimation, rather than data size, taxon sampling, or methods for tree reconstruction. Although glaucophytes were excluded from our analyses due to a shortage of sequence data, we found that the recovery of a sister–group relationship between red algae and green plants primarily depends on gene sampling in phylogenetic inferences from <10,000 aa positions. Phylogenetic analyses of data sets with fewer than 10,000 aa positions, which can be prepared without large-scale sequencing (e.g., expressed sequence tag analyses), are practical in challenging various unresolved issues in eukaryotic evolution. However, our results indicate that severe biases can arise from gene sampling in multi-gene inferences from <10,000 aa positions. We also address the validity of fast-evolving gene exclusion in multi-gene phylogenetic analyses, in light of this gene sampling bias.

Introduction

Resolving the global phylogeny of eukaryotes is one of the most fundamental questions in biology that has been addressed using molecular phylogenetic techniques. Since the last century, small subunits of ribosomal RNA (SSU rRNA) sequences have been sampled from phylogenetically diverged eukaryotes and extensively analyzed to trace organismal evolution. The SSU rRNA gene is indeed powerful in identifying evolutionary “relatives” of the organism(s) of interest and, as a result, has been considered as one of the key markers for eukaryotic phylogeny (e.g., Van de Peer et al. 1993). However, the phylogenetic information in SSU rRNA genes (or any single gene) is generally insufficient to resolve the organismal relationships among distantly related lineages/groups with high statistical support. In theory, this difficulty in phylogenies based on single-gene data sets (single-gene phylogenies) can be overcome by phylogenetic analyses of data sets comprised of multiple nucleic-encoded genes (multi-gene phylogenies), because phylogenetic information in multiple genes should be greater than that present in any single gene. As anticipated, multi-gene analyses conducted to date have successfully provided novel insights into deep splits in eukaryotic evolution (e.g., Burki et al. 2008).

The current major working hypothesis for the global eukaryotic phylogeny assumes that most eukaryotic organisms can be classified into six putative “supergroups” (Simpson and Roger 2004; Adl et al. 2005). One of the six putative supergroups, Plantae (or Archaeplastida), is comprised of glaucophytes, red algae, and green plants (green algae plus land plants) all of which possess “primary” plastids that are direct descendants of an endosymbiotic cyanobacterium (Adl et al. 2005). It is generally thought that plastid acquisition via cyanobacterium–eukaryote endosymbiosis is a rare evolutionary event, because this process requires drastic modifications of both host and endosymbiotic genomes, as well as inventions of host cellular machineries to enslave the endosymbiont as an organelle (e.g., the protein targeting machinery). This argument therefore favors Plantae monophyly, which invokes a single cyanobacterium–eukaryote endosymbiosis in a common ancestor of the members of this group.

Although the monophyly of Plantae is straightforward and well accepted, it is rather difficult to confirm by nucleus-encoded gene phylogenies. Neither Plantae monophyly nor the sisterhood between red algae and green plants were recovered in analyses of any single-gene data sets (e.g., Stiller et al. 2001), except that of translation elongation factor 2 (ef2) by Moreira et al. (2000) that reconstructed a robust clade of red algae and green plants (R + G monophyly). However, Kim and Graham (2008) recently demonstrated that ef2 sequences from cryptomonads, haptophytes, and the katablepharid Leucocryptos marina showed robust affinities to the red algal and green plant homologues, indicating that the phylogeny of ef2 does not exclusively unite red algae and green plants. On the other hand, “phylogenomic” analyses based on multiple genes—typically those including >30,000 amino acid (aa) alignment positions—have successfully recovered the Plantae clade (Rodríguez-Ezepeleta et al. 2005; Burki et al. 2007; Patron et al. 2007; Rodríguez-Ezepeleta, Brinkman, Burger, et al. 2007; Burki et al. 2008). It is noteworthy that the statistical support for the monophyly of Plantae was low to medium in analyses including haptophytes.
and cryptomonads (Burki et al. 2007; Patron et al. 2007; Rodrı́guez-Ezpeleta, Brinkman, Burger, et al. 2007; Burki et al. 2008), in sharp contrast to the Plantae clade recovered with high statistical support in Rodrı́guez-Ezpeleta et al. (2005) including neither of these two groups.

Smaller multi-gene phylogenetic analyses of 10,000 aa positions, many of which are based on fewer than 10 genes, have often been applied to issues in eukaryotic evolution that cannot be solved by single-gene analyses (Baldauf et al. 2000; Arisue, Hashimoto, Lee, et al. 2002; Arisue, Hashimoto, Yoshikawa, et al. 2002; Nozaki et al. 2003; Hampl et al. 2005; Harper et al. 2005; Nozaki 2005; Kim et al. 2006; Simpson et al. 2006; Sakaguchi et al. 2007; Simpson et al. 2008). However, the phylogenetic positions of glaucophytes, red algae, and green algae (or the relationship between red algae and green plants) inferred from multi-gene data sets including 10,000 alignment positions appeared to be unstable. For instance, Plantae/R + G monophyly was robustly supported in an analysis of 6,735 amino acid (aa) positions (Hackett et al. 2007), whereas Nozaki et al. (2007) contested Plantae/R + G monophyly based on an analysis of 5,216 aa positions.

In order to clarify the precise reason(s) for the instability of Plantae/R + G monophyly in the two previous multi-gene studies (Hackett et al. 2007; Nozaki et al. 2007), we systematically evaluated the significance of R + G monophyly by analyzing multi-gene data sets varying in size from ~2,000 to ~10,000 aa positions. Importantly, gene sampling in multi-gene data sets appeared to largely affect the inference of the relationship between red algae and green plants. We often observed that analyses of two multi-gene data sets comprised of different gene sets reached opposite conclusions regarding the relationship between red algae and green plants, although the two data sets were similar to one another in data size. In addition, our analyses clearly showed that the signal for the monophyly of red algae and green plants in the ef2 gene was extraordinarily stronger than those in other genes. Based on the results presented here, we discuss the potential sources of the incongruity between the phylogenetic positions of glaucophytes, red algae, and green algae in Hackett et al. (2007) and those in Nozaki et al. (2007). Finally, we argue that the systematic exclusion of fast-evolving genes from multi-gene data sets enhances the artifactual impact of gene sampling.

### Materials and Methods

#### Single-Gene Data Sets

We prepared 27 single-gene data sets for multi-gene phylogeny. The details of the individual single-gene data sets are summarized in table 1. Among these data sets, 22 were originally generated in Arisue et al. (2005) and used in this study after excluding closely related sequences. We constructed five additional data sets consisting of...
eukaryotic release factor 1 (erf1), eukaryotic release factor 3 (erf3), vacuolar ATPase A subunit (vatpa), vacuolar ATPase B subunit (vatpb), and Dom34 endonuclease (dom). The erf1, erf3, and vatpb alignments used in Inagaki et al. (2002, 2003) and Arisue, Hashimoto, Yoshikawa, et al. (2002), respectively, were updated by adding new sequences. vatpa and dom sequences retrieved from GenBank database were manually aligned. Only unambiguously aligned positions were subjected to phylogenetic analyses. Our preliminary analyses suggested that neither paralogous nor xenologous genes existed in the 27 single-gene data sets (data not shown). In this study, we only evaluated the relationship between red algae and green plants because few glaucophyte sequences are available for the above 27 genes. For the same reasons, haptophytes, cryptomonads, and rhizarians were also excluded. It should be mentioned, however, that the absence of haptophytes and cryptomonads in our data sets may, to some extent, bias the credibility of R + G monophyly. In recent multi-gene phylogenies (e.g., Patron et al. 2007; Burki et al. 2008), haptophytes and cryptomonads showed an affinity with members of Plantae. If haptophytes and cryptomonads are genuinely nested in the Plantae clade in the host phylogeny, analyses excluding the two groups are likely to give an artifactually high support for R + G monophyly. Nevertheless, the principal subject in this study—the impact of gene sampling on the significance of R + G monophyly (see below)—is testable by analyses excluding cryptomonads and haptophytes.

Multi-Gene Data Sets

Each single-gene data set contained representative homologues from members of seven putative monophyletic assemblages in eukaryotes, namely 1) Opisthokonta, 2) Amoeboza, 3) red algae, 4) green plants, 5) a grouping of alveolates and stramenopiles (SA clade), 6) a grouping of diplomonads and parabasalids (DP clade), and 7) a grouping of jakobids, euglenozoans, and heterolobosea (JEH clade) (see table 1). The internal relationship in each group was constrained in advance (supplementary table S1, Supplementary Material online). Previously published multi-gene phylogenies have reliably reconstructed a robust monophyletic clade in multi-gene phylogenies (Arisue et al. 2005; Simpson et al. 2006). The DP clade was further supported by “archaeal” prolyl-tRNA synthetase genes exclusively found in the two lineages (Anderson et al. 2005). Jakobids, euglenozoans, and heteroloboseans formed a robust monophyletic clade in multi-gene phylogenies (Simpson et al. 2006; Rodríguez-Ezepeleta, Brinkman, Burger, et al. 2007). This grouping is further consistent with a characteristic aa insertion in ribosomal protein L24A (Rodríguez-Ezepeleta, Brinkman, Burger, et al. 2007).

We combined all of the 27 single-gene data sets to generate a “27g” multi-gene data set. In theory, the phylogenetic signals in the 27 genes should not be significantly incongruent with one another. If this assumption is violated, the multi-gene inference as a whole could be critically misled. We therefore examined whether this assumption was valid for the 27 single-gene data sets prior to multi-gene analyses. As in Iida et al. (2007), the phylogenetic “signal” in each gene was represented by a set of P values for the Shimodaira–Hasegawa (SH) test (Shimodaira 2002) calculated over 945 test trees that are possible for the 7 monophyletic groups (see below). Subsequently, the 27 sets of the SH P values were subjected to hierarchical clustering to identify potentially problematic data sets for multi-gene phylogeny. The clustering suggested that the signals of ef2 and α-tubulin (tba) genes appeared to be largely different from those of other genes. The two genes were reciprocally excluded to generate “ef2” and “tba” multi-gene data sets. Finally, the “ef2, tba” data set was prepared by excluding both ef2 and tba genes from the 27g data set. The four multi-gene data sets were then subjected to maximum likelihood (ML) analyses (see below).

To investigate the impact of gene sampling on R + G monophyly, 50 of each “ef2,” “10g,” “15g,” “20g,” and “25g” multi-gene data sets were prepared by combining 5, 10, 15, and 20 randomly sampled genes from the ef2 multi-gene data set. We excluded ef2 gene, which has an exceptionally strong signal for R + G monophyly (see below), from the above data sets. In addition, a single gene was reciprocally excluded from the ef2 data set to generate 25 of “25g” multi-gene data sets.

We generated two additional series of multi-gene data sets using the procedure described above: 1) 50 of each “5g,” “10g,” “15g,” “20g,” and “25g” data sets that included ef2 gene and 4, 9, 14, and 19 genes randomly sampled from the ef2 multi-gene data set, and 2) 50 of each “5g,” “10g,” “15g,” “20g,” and “25g” data sets that included tba gene and 4, 9, 14, and 19 genes randomly sampled from the ef2 multi-gene data set. Note that the last series of multi-gene data sets excluded the ef2 gene data set.

Multi-Gene Phylogenetic Analyses

Generally, a multi-gene data set can be analyzed under two different models. In the analysis under the “concatenate” model conditions (concatenate analyses), a multi-gene data set is treated as a large single-gene data set, and parameters are optimized for the entire alignment. In contrast, in the analysis under the “separate” model conditions, parameters are optimized for each gene in the multi-gene alignment, taking into account the potential heterogeneity of evolutionary tempo and mode across genes. We examined the two models using a data set comprised of the 27 genes listed in table 1, and found that the separate model was a better approximation for the data set than the concatenate model. The details of the comparison are presented in supplementary figure S1, Supplementary Material online. Thus, all multi-gene data sets were analyzed under the separate model conditions.

We subjected the 27g, ef2, tba, and ef2, tba data sets to ML analyses under the separate model conditions. The four data sets included homologues of seven major
eukaryotic groups (see above). We assumed the monophy- 
lies of the above eukaryotic groups in advance, and exhaustively searched for the ML tree from 945 possible test trees 
for the seven groups. Log-likelihoods (lnLs) of the test trees 
were calculated based on each single-gene data set with op-
timization of branch lengths and a parameter for among-site 
rate variation (ASRV). Subsequently, for each test tree, the 
“total” lnL value was obtained by summing the lnL values 
calculated from 25 to 27 single-gene data sets. The tree with 
the highest total lnL value was thus the ML tree. Amino acid 
substitutions in the data were modeled under the WAG 
matrix with ASRV approximating a discrete gamma (Γ) 
distribution with four categories (WAG + Γ).

To assess the robustness of the ML trees, bootstrap 
(BP) analyses were carried out. We generated 100 BP rep-
licates from each single-gene data set by SEQBOOT in 
PHYLIP v.3.6 (Felsenstein 1993), and then these “sing-
gle-gene” BP replicates were combined into 100 “multi-
gene” BP replicates. The tree search under the separate 
WAG + Γ model was repeated for each multi-gene BP rep-
licate as described above. The resultant BP trees were sub-
jected to CONSENSE in PHYLIP (Felsenstein 1993) to 
obtain ML BP values. We modified TREE-PUZZLE 
v.5.2 (Schmidt et al. 2002) to calculate the lnLs of user-de-
defined trees in parallel and utilized this program for the anal-
yses described above.

Analyses of the Data Sets Comprised of Randomly 
Sampled Genes

Based on the 25g$$\text{ef}$$ multi-gene data set and those 
comprised of 5, 10, 15, and 20 randomly sampled genes, 
the 945 test trees were scored under the separate WAG 
+ Γ model as described above. The test trees can be split 
into two types: 1) 105 “R + G” trees with the R + G clade 
and 2) 840 “R × G” trees in which red algae and green 
plants were separated. For each data set, we selected the 
best R + G tree and the best R × G tree based on the 
lnL values calculated under the separate model conditions. 
The difference in the lnL of the two trees (ΔlnL) was 
calculated by subtracting the lnL value of the best R × G tree 
from that of the best R + G tree. A positive (or negative) 
ΔlnL indicates the degree of dominance of the R + G tree 
(or R × G tree) over the other.

Support values for R + G monophyly were also ob-
tained from the above multi-gene data sets using the re-
sampling estimated log-likelihood (RELL) approxima-
tion. Sitewise lnL values were calculated for each 
single-gene data set under the WAG + Γ model using 
TREE-PUZZLE. Then, these sitewise lnL data were ana-
alyzed with the TOTALML program in the MOLPHY v.2.3 
package (Adachi and Hasegawa 1996) for calculation of 
RELL values.

Multi-gene Analyses Excluding “Fast-Evolving” Genes

We defined the average pairwise distance between 
taxa as the overall evolutionary speed of a particular sin-
gle-gene data set. The 27 genes were sorted according to 
their average pairwise distances calculated under the 
WAG + Γ model by Tree-Puzzle. The “20g$$\text{ef}$$,” 
“15g$$\text{ef}$$,” and “10g$$\text{ef}$$” data sets were generated from the 
$$\Delta$$ef2 multi-gene data set by excluding 6, 11, and 16 fast-
est-evolving genes, respectively. To evaluate the impact of 
fast-gene exclusion, the ef2 gene with an intense signal (see 
below) was excluded. The ΔlnL scores and RELL values 
for R + G monophyly were then calculated for the three 
multi-gene data sets (see the previous section for details).

Results and Discussion

Multi-Gene Analyses Recovered the Monophyly of Red 
Algae and Green Plants

In the phylogenetic analysis of the 27g data set (11,164 
aa positions in total), the R + G clade was recovered with 
a BP of 71% (fig. 1A). In the hierarchical clustering of the 
$SH P$ values shown in figure 1B, ef2 and tba genes were 
separated from the remaining 25 genes by a large Euclidean 
distance, suggesting that the two gene signals potentially 
conflict with those of the other genes. We could find no 
obvious correlation between the $P$ value–based clustering 
and data size (fig. 1B). In the ef2 gene–based SH test, 
all 105 R + G trees and four R × G trees received $P > 0.05$, suggesting that the ef2 gene strongly preferred R 
+ G monophyly (data not shown). Similarly, the tba gene 
appeared to prefer a grouping of Opisthokonta and the DP 
clade (Op + DP grouping); One hundred and five of 135 
trees with $P > 0.05$ showed the Op + DP grouping (data 
not shown). This tba-specific signal is most likely attribut-
able to a putative lateral transfer of tba gene recently pro-
posed by Simpson et al. (2008). According to the $SH P$ 
values, all genes except ef2 and tba genes did not show a 
strong preference for any particular tree (see the supplemen-
tary data, Supplementary Material online).

Importantly, R + G monophyly was consistently re-
constructed in the ML trees in both analyses of the $\Delta$ef2 
and $\Delta$tba data sets (10,493 and 10,793 aa positions in total, 
respectively), although the corresponding BP value was 
dramatically different for the two analyses (fig. 1A). R + G 
monophyly received a BP of 91% in the $\Delta$tba analysis,
whereas the same relationship was less strongly supported 
in the $\Delta$ef2 analysis (BP = 47%). Consistent with the ef2 
gene–based SH test (see above), the robustness of R + G 
monophyly in the 27g analysis was largely attributable to 
the presence of the ef2 gene. A large contribution of the ef2 
gene in the strength of R + G monophyly in multi-gene 
phylogenies was also detected in a single-gene jackknife 
analysis (Sakaguchi et al. 2007). On the other hand, there 
was no apparent sign that the tba gene positively contrib-
uted to R + G monophyly. This is also consistent with the 
result from the SH test in which the tba gene signal was 
unrelated to R + G monophyly (see above). In a multi-gene 
BP analysis that excluded both ef2 and tba genes (25 genes 
with 10,122 aa positions in total), R + G monophyly received a BP of 57% (fig. 1A).

Although the precise reason for the extraordinary signal of 
ef2 gene for R + G monophyly remains unclear, it is 
worthwhile considering the possible recombination event be-
tween red algal and green plant ef2 genes suggested by Stiller 
et al. (2001). They proposed that the region corresponding to
Fig. 1.—Multi-gene phylogenetic analyses under the separate model conditions. (A) The ML tree inferred from the 27g multi-gene data set. The monophyllies of seven groups were constrained: 1) green plants, 2) red algae, 3) stramenopiles plus alveolates (SA), 4) jakobids plus euglenozoans plus heterolobosea (JEH), 5) Amoebozoa, 6) diplomonads plus parabasalids (DP), and 7) Opisthokonta. The ML tree was exhaustively searched for among 945 possible test trees for the seven monophyletic groups. Multi-taxon groups are shown as triangles. Heights of the triangles are proportional to the average numbers of taxa included in each single-gene data set. Branch lengths and widths of triangles are proportional to the average branch lengths of the 27 genes (weighted by the size of each gene). BP values based on the 27g data set are presented on the corresponding edges. On the edge uniting red algae and green plants, the BP value for the ef2, Afta, and def2 data sets are presented. (B) Hierarchical clustering of the SH P values calculated from the 27 single-gene data sets. Abbreviations for gene names are shown in table 1. The values from the def2, Afta, and def2 data sets are presented. (C) Heat map presentation of SH P values calculated were are from the 27 single-gene data sets. The significance of R dependent on Plantae + G monophyly was examined by multi-gene analyses of randomly sampled genes. Figures 2A–C confirm that 1) the data size was positively correlated with the number of genes considered and 2) the data sets at each node point were similar to one another in size. We firstly examined the multi-gene data sets of randomly sampled non-def2 genes (fig. 2D and G). The ef2 gene, whose intense signal seemed incongruent with the signals of other genes (see above), was excluded from the first set of analyses. In the 5g-def2 analysis, all MinL values calculated were the putative recombination region (“nonrecombination” positions) failed to recover R + G monophyly, contrasting sharply to the highly supported R + G monophyly recovered in the same MP analysis of the putative recombination region (Stiller et al. 2001). Curiously, using the ML method, we recovered R + G monophyly from the nonrecombination positions in our ef2 data set (supplementary fig. S2, Supplementary Material online). Our analyses suggest that the tree topology inferred from the nonrecombination positions is simply sensitive to the methods used for tree reconstruction (MP vs. ML). Although it is still difficult to completely refute the putative ef2 gene recombination, the MP analyses of partial ef2 alignments in Stiller et al. (2001) cannot be considered as strong evidence to support the gene recombination scenario for the evolution of the ef2 gene in red algae and green plants.

The Relationship between Red Algae and Green Plants Depends on Gene Sampling

Rodriíguez-Ezpeleta, Brinkman, Roure, et al. (2007) indicated that equal or larger than 20,000 aa positions were required to support the monophyly of Plantae with BP >90% by a series of phylogenetic analyses of randomly sampled alignment positions from a master data set, including larger than 30,000 aa positions. This pioneering work indicated that combining the signals from a large number of alignment positions is essential to suppress stochastic errors in sequence evolution (i.e., nonphylogenetic noise) and recover ancient splits with high statistical support. However, it is noteworthy that no haptophyte or cryptomonad sequence was considered in the analyses presented in Rodríguez-Ezpeleta, Brinkman, Roure, et al. (2007). When haptophytes and cryptomonads are included, the data size required to recover Plantae monophyly with high statistical support is almost certainly larger than that estimated from analyses excluding the two groups. In practice, however, generating massive sequence data from key taxa to resolve global eukaryotic phylogeny is not an easy (or inexpensive) task, and to date, multi-gene phylogenies based on >30,000 aa positions remain fairly rare. Consequently, it is important to clarify the potential factor(s) that can bias phylogenetic inferences from smaller multi-gene alignments with fewer than 10,000 aa positions (i.e., typically those comprised of fewer than 10 genes). In this study, we have specifically investigated the support for Plantae/R + G monophyly in inferences from ~2,000 to ~10,000 aa positions. However, because we generally combine genes as primary “units” when preparing a multi-gene data set, we analyzed data sets comprised of randomly sampled genes, instead of focusing on random samplings of alignment positions.

The significance of R + G monophyly was examined by multi-gene analyses of randomly sampled genes. Figures 2A–C confirm that 1) the data size was positively correlated with the number of genes considered and 2) the data sets at each node point were similar to one another in size. We firstly examined the multi-gene data sets of randomly sampled non-def2 genes (fig. 2D and G). The ef2 gene, whose intense signal seemed incongruent with the signals of other genes (see above), was excluded from the first set of analyses. In the 5g-def2 analysis, all MinL values calculated were...
negative, suggesting that the R + G tree (or R + G monophyly) was not positively supported in these small data sets (fig. 2D). In the 10g_{ef2}–20g_{ef2} analyses, both negative and positive ΔInL values were detected (fig. 2D). The overall distribution of the ΔInL values (fig. 2D) and that of the RELL values for R + G monophyly (fig. 2G) appeared to be shifted upward by the progressive inclusion of genes (fig. 2A). In the analyses of the 5g_{+ef2}–20g_{+ef2} data sets including the ef2 gene, R + G trees appeared to dominate over R × G trees; regardless of the number of genes in the multi-gene data sets, all ΔInL values were positive (fig. 2E) and R + G monophyly received large RELL values (fig. 2H). Inclusion of the tba gene in the multi-gene data sets had no large impact on the ΔInL values (fig. 2F), because the tba-specific signal exclusively affects the relationship between Opisthokonta and the DP clade (see above).
Thus, we did not conduct RELL calculations based on the 5g\textsuperscript{tba} - 20g\textsuperscript{tba} data sets.

The trends observed in the 5g\textsuperscript{ef2} - 20g\textsuperscript{ef2} analyses can be rationalized as follows. Because the signal for the R + G trees was faint in each gene, the signal for R + G monophyly was masked by nonphylogenetic noise in the analyses of data sets comprised of a small number of genes. Overall, the signal for R + G monophyly was strengthened and finally surpassed the competing nonphylogenetic signals in analyses of large numbers of genes. Nevertheless, both positive and negative ΔlnL values were observed in the 10g\textsuperscript{ef2} - 25g\textsuperscript{ef2} analyses (fig. 2D), suggesting that the inference for the relationship between red algae and green plants primarily depended on gene sampling. It is also worth mentioning that the intense signal of ef2 gene completely canceled the nonphylogenetic noise from gene sampling in the 5g\textsuperscript{ef2} - 20g\textsuperscript{ef2} analyses (fig. 2E). Unfortunately, it is unclear whether the ef2 signal is truly in conflict with those of other genes. For a cautious assessment of the relationship between red algae and green plants, the results of the two separated multi-gene analyses, one with the ef2 gene and the other without, should be carefully evaluated.

Although we analyzed no data sets with ≥25 randomly sampled genes in the current study, the plots (fig. 2D) imply that the dependency of R + G monophyly on gene sampling most likely diminishes, and eventually disappears, in analyses considering a much larger number of genes.

Perspectives on Nozaki et al. (2007) and Hackett et al. (2007)

Recent large-scale multi-gene inferences (Burki et al. 2007; Rodriguez-Ezpeleta, Brinkman, Burger, et al. 2007; Burki et al. 2008) consistently recovered Plantae/R + G monophyly, but the support values obtained were low to moderate in these analyses, despite the extremely large number of alignment positions considered. Here, we demonstrated that, in multi-gene analyses of <10,000 aa positions, the relationship between red algae and green plants (and probably the relationship among glaucophytes, red algae, and green plants as well) is highly dependent on gene sampling. Indeed, two recent studies based on multi-gene data sets with fewer than 10,000 aa positions reached contradictory conclusions regarding the relationship among glaucophytes, red algae, and green plants. Hackett et al. (2007) analyzed a “16-gene” data set (6,735 aa positions; ef2 gene included) and successfully recovered Plantae/R + G monophyly with high support values, whereas this group was not recovered in the analyses of a “19-gene” data set (5,216 aa positions; no ef2 gene included) by Nozaki et al. (2007). The presence or absence of the ef2 signal favoring R + G monophyly (see fig. 2E and H) may be a primary cause of the incongruity regarding Plantae/R + G monophyly. In addition, the difference in (non-ef2) gene sampling between the two multi-gene data sets of these studies could have further magnified the incongruity as there were only five genes in common to the data sets. Although it requires more time and effort to assemble the data sets, Plantae/R + G monophyly should be examined by analyzing multi-gene data sets that are free from the bias from gene sampling (e.g., those including >30,000 aa positions).

The only other factor that is likely to contribute to the incongruities between the inferences in Hackett et al. (2007) and Nozaki et al. (2007) is the difference in taxonomic sampling of the two studies.

Fast-evolving genes have likely accumulated multiple substitutions and/or significant compositional biases that cannot be adequately accounted for under the current widely used models for phylogenetic estimation. Thus, it may be sensible to systematically exclude fast-evolving genes from multi-gene phylogenies as a means to ameliorate potential artifacts in tree reconstruction (Brinkmann et al. 2005; Philippe et al. 2005; Nishihara et al. 2007). Nevertheless, this may not work properly for multi-gene data sets that are biased by gene sampling. Because bias due to gene sampling is not necessarily attributable to fast-evolving genes, the gene exclusion procedure may not improve the overall multi-gene inferences. Indeed, when the ΔlnL and RELL values from the 10g\textsuperscript{EF} - 20g\textsuperscript{EF} multi-gene data sets were compared with the corresponding values from the 10g\textsuperscript{ef2} - 25g\textsuperscript{ef2} multi-gene data sets (filled squares in fig. 2D and G), the exclusion of fast-evolving genes did not impact on the relationship between red algae and green plants. It is also important that the data size after gene exclusion ought to be sufficiently large so as not to increase the impact of gene sampling that is masked prior to gene exclusion.

When the “postgene exclusion” data sets include fewer than 10,000 aa positions, it is important to be prepared for a severe impact of gene sampling on inferences of the relationship between red algae and green plants. Considering the connection between gene exclusion and the bias from gene sampling, the distant relationship between red algae and green plants (and the Plantae paraphyly) recovered in Nozaki et al. (2007) should be treated with caution, because this particular relationship was inferred from a post-gene exclusion data set with 5,216 aa positions.

Acknowledgments

We thank Drs. M. Hasegawa (Fudan University, China) and K. Takishita (Japan Agency for Marine-Earth Science and Technology, Japan) for critical reading of this manuscript. This research was partially carried out by using the PACS-CS Cluster in the Center for Computational Sciences, University of Tsukuba.

Literature Cited


Arisue N, Hasegawa M, Hashimoto T. 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood