J Mol Evol (2003) 56:330–340
DOI: 10.1007/s00239-002-2404-3

JOURNAL OF MOLECULAR EVOLUTION

Probabilistic Analysis Indicates Discordant Gene Trees in Chloroplast Evolution

Claus Vogl,[1] Jonathan Badger,[2] Paul Kearney,[2] Ming Li,[3] Michael Clegg,[4] Tao Jiang[1]

[1] Department of Computer Science, University of California, Riverside, CA 92521, USA
[2] Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[3] Department of Computer Science, University of California, Santa Barbara, CA 93106, USA
[4] Department of Botany and Plant Sciences, University of California, Riverside, CA 92521-0124, USA

Received: 29 August 2001 / Accepted: 17 October 2002

Abstract. Analyses of whole-genome data often reveal that some genes have evolutionary histories that diverge from the majority phylogeny estimated for the entire genome. We present a probabilistic model that deals with heterogeneity among gene trees, implement it via the Gibbs sampler, and apply it to the plastid genome. Plastids and their genomes are transmitted as a single block without recombination, hence homogeneity among gene trees within this genome is expected. Nevertheless, previous work has revealed clear heterogeneity among plastid genes (e.g., Delwiche and Palmer 1996). Other studies, using whole plastid genomes of various algae and land plants, found little additional heterogeneity (Martin et al. 1998; Adachi et al. 2000). We augment the earlier studies by using a data set of 14 taxa: 6 land plants, 2 green algae, a diatom, 2 red algae and a cryptophyte, the cyanelle of the glaucocystophyte *Cyanophora*, and the blue–green alga *Synechocystis* as an outgroup. Contrary to the earlier analyses, we cannot find even a single, dominant consensus tree. Therefore, we formulate a probabilistic model that divides the genes into two sets: those that follow the consensus tree and those that have independent gene trees. No particular tree is supported by more than three-fourths of the genes. But the set of genes that follows a certain tree is fairly independent of data processing and the method of analysis. With one possible exception, we find no evidence for collinear or functionally related genes to follow similar trees. The phylogenetic pattern also seems independent of bias in amino acid composition. Among possible explanations for the observed phenomenon, the hypothesis that different genes have different covarion structures is difficult to assess. But gene duplication may be possible through the inverted or direct repeat regions, while horizontal gene transfer seems less likely. In contrast to green algae and land plants, inverted repeat regions in red algae and in *Cyanophora* show abundant differences among the copies. Thus, genes may get duplicated when they are recruited into the inverted repeat region and one of the two copies may be lost after leaving the inverted repeat region.

Key words: Chloroplast — Phylogeny — Gene duplication — Probabilistic model — Gibbs sampling

Correspondence to: Claus Vogl, Institute of Animal Husbandry and Genetics, University of Veterinary Medicine, Vienna, Veterinärplatz 1, A-1210 Vienna, Austria; *email:* vogl@i122server.vu-wien.ac.at

Introduction

The comparative analysis of whole-genome sequence data is emerging as a major scientific enterprise. For example, much is being revealed about the minimal gene content of various eubacterial lineages through comparative analyses (Doolittle, 1999; Ochman and Moran 2001). Moreover, the comparative analysis of whole genomes provides the most complete infor-

mation on organismic history and it reveals the evolutionary pathways followed by the various genes that comprise a genome. Often gene histories are highly divergent within a genome, such that the notion of an organismic lineage may be rendered meaningless, because early prokaryotic genomes represent a composite of genes with different histories (Doolittle 1999). Likewise gene trees estimated from the eukaryotic lineage that led to algae and land plants may be heterogeneous because of the assimilation of genes with a prokaryotic origin into the eukaryotic genome.

Inference of gene phylogenies provides a powerful tool for the identification of genes with divergent histories. But given a choice of divergent phylogenies, what methods are available for choosing a consensus tree? How are minority trees best evaluated? and How many separate evolutionary pathways are represented within a genome? In this article we seek to address these questions by presenting a probabilistic model that partitions gene histories into consensus and minority trees and calculates the probability that particular genes conform to various trees. We use the plastid genome as a model because a number of complete sequences are available and because of its relatively small and compact structure.

The evolutionary origins of the plastid genome are well established. According to the now well-accepted serial endosymbiosis theory (e.g., Melkonian, 1996), eukaryotic hosts phagocytized and retained formerly photosynthesizing prokaryotes. Subsequently, these prokaryotic endocytobionts underwent genetic reduction involving transfer of some chloroplast genes to the eukaryotic nucleus and loss of others and were transformed into plastids. Most likely, there was a single primary endocytobiosis event, i.e., all plastids are a monophyletic group derived from a single free-living blue–green alga. There is, however, ample evidence for secondary endocytobiosis, where a eukaryotic host took up a eukaryotic alga. The primary endocytobiosis event must have been ancient, as fossils of red algae have been found in strata that date to about 1200 MYA (Butterfield 2000). Most multicellular algae (e.g., brown and green algae) were abundant 1000 MYA (Knoll 2000). This means that the secondary endocytobiosis that gave rise to the brown (stramenopile) algal plastid must have occurred before 1000 MYA.

In their eukaryotic hosts, plastids are inherited as a unit without recombination, such that all genes in the chloroplasts should have the same phylogeny. Despite this complete linkage in transmission, heterogeneity among genes in estimated phylogenies is commonly observed (see below). There are several reasons to expect such heterogeneity. First, with such ancient divergence times, saturation of changes may compromise the phylogenetic signal. Second, horizontal gene transfer (HGT) or gene duplication and subsequent loss, i.e., paralogy, may lead to inconsistency in gene phylogenies. While HGT and paralogy are quite different phenomena, they produce similar deviations from a common phylogenetic history in a phylogenetic analysis (e.g., Eisen, 2000). Finally, varying patterns of codon utilization and other selective regimes can lead to large deviations among genes when accumulated over such long periods of time. Lockhart et al. (1998) discuss this phenomenon under the term "covarion structure of genes".

Hence phylogeny needs to be reconstructed carefully to separate signal from noise. Quality of reconstruction is affected by phylogenetic analysis tools (see Materials and Methods). Increased resolution also comes from addition of taxa or of genes or of both. Insufficient taxon sampling is often cited as a major source of error in phylogenetic analysis. A recent computer simulation study, however, shows increasing sequence information to improve phylogeny reconstruction more efficiently than increasing the number of taxa sampled (Rosenberg and Kumar 2001).

For some genes, sequence information from many taxa is available. Thus, a phylogenetic analysis of the *rbcL* gene indicated that the *rbcL* genes of red and brown algae are more closely related to each other and to some α- and β-proteobacteria than to blue–green algae and the other chloroplasts used in that study (Delwiche and Palmer 1996). Martin and Schnarrenberger (1997) explained this phylogenetic pattern of *rbcL* by functional redundancy of the gene in the prokaryotic ancestors of chloroplasts and subsequent loss during the reduction to the plastid genome, i.e., by paralogy. In an analysis of the *secA* gene, Valentin (1997) found a similarly deep split between red and brown algae and land-plant chloroplasts and argued therefore for a polyphyletic origin of chloroplasts. In a reanalysis of the same data, Barbrook et al. (1998) found the support for polyphyly of the chloroplasts decreasing if unequal AT content was accounted for by LogDet transformation of distances (Lockhart et al. 1994).

For some organisms, all chloroplast sequence information is available. Larger-scale phylogenetic analyses using all information from chloroplast protein encoding genes (Martin et al. 1998; Adachi et al. 2000) of land plants, *Euglena*, a red alga and a diatom, the cyanelle of *Cyanophora*, and *Synechocystis* as an outgroup found little evidence for HGT or paralogy. In the first of these studies (Martin et al. 1998), a single consensus tree was identified using a wide array of data processing and phylogenetic methods. The most likely gene trees were identical to the consensus tree or at least did not reject it. Only the fast-evolving *rpoB*, *rpoC1*, and *rpoC2* genes as well as *rps8* rejected the consensus tree at the 0.95

significance level. Lockhart et al. (1999) challenged these results, claiming that the placement of *Odontella* with *Porphyra* may be the result of a shared bias in amino acid composition. In a careful reanalysis of the same data without *rbcL* (Adachi et al. 2000), four possible consensus trees instead of just one were found. One of these trees (Tree 2), which separates *Odontella* from *Porphyra*, received little support; the other three differed only in the position of where the outgroup *Synechocystis* joins the ingroup, i.e., in the position of the root of the chloroplast phylogeny. No attempt was made to detect HGT or paralogy in the second analysis.

Shared gene arrangements provide independent evidence for phylogenetic relationships. The conservation of gene arrangements is striking in the chloroplasts of the cryptophyte *Guillardia* and the red alga *Porphyra*, suggesting a close phylogenetic relationship between these two chloroplasts (Douglas and Penny 1999). Both also share many arrangements with the red alga *Cyanidium* (Glöckner et al. 2000) and, to a lesser degree, with the diatom *Odontella* (Turmel et al. 1999). These shared arrangements are indicative of monophyly of these four taxa and thus independently validate the study of Adachi et al. (2000).

In this study, we augment the chloroplast genome data set to 14 taxa: 6 land plants (*Oryza sativa, Zea mays*, *Nicotiana tabacum*, *Arabidopsis thaliana*, *Pinus thunbergii, Marchantia polymorpha*), 2 green algae (*Nephroselmis olivacea*, *Chlorella vulgaris*), a diatom (*Odontella sinensis*), 2 red algae (*Cyanidium caldarium*, *Porphyra purpurea*), the cryptophyte *Guillardia theta*, the cyanelle of the glaucocystophyte *Cyanophora paradoxa*, and the homologous genes of the blue–green alga *Synechocystis sp*. as an outgroup. A set of 48 genes was shared among these taxa. But we include neither *Euglena gracilis* nor nonphotosynthesizing taxa with plastids because that would have reduced the number of genes available. Furthermore, *Euglena* has some rather unusual genetic characteristics, such as introns and differences in codon usage (Morton 1998) and amino acid composition (Adachi et al. 2000). This increase in number of both taxa and genes should increase accuracy in phylogenetic reconstruction.

Our primary goal is to identify congruence or incongruence among genes in their phylogenetic pattern, rather than to infer a ''best'' phylogeny. Specifically, we identify subsets of genes exhibiting the same phylogeny (or set of phylogenies), as opposed to others exhibiting varying phylogenies. To this end, we introduce a probabilistic model to infer the (approximate) posterior probability of consensus trees and the probability of the different genes to conform to the consensus tree or to different trees. The model is implemented via a Gibbs sampler (Casella and George 1992; Gelman et al. 1995).

Table 1. Abbreviations of taxa

Species name	Abbreviation
Oryza sativa	ory
Zea mays	zea
Nicotiana tabacum	nic
Arabidopsis thaliana	ara
Pinus thunbergii	pin
Marchantia polymorpha	mar
Nephroselmis olivacea	nep
Chlorella vulgaris	chl
Odontella sinensis	odo
Cyanidium caldarium	cyc
Porphyra purpurea	por
Guillardia theta	gui
Cyanophora paradoxa	cyp
Synechocystis sp.	syn

Materials and Methods

We used the complete cpDNA sequences of the land plants *Oryza sativa* (NCBI database accession number NC 001320), *Zea mays* (NC 001666), *Nicotiana tabacum* (NC 001879), *Arabidopsis thaliana* (NC 000932), *Pinus thunbergii* (NC 001631), *Marchantia polymorpha* (NC 001319), the two green algae *Nephroselmis olivacea* (NC 000927) and *Chlorella vulgaris* (NC 001865), the diatom *Odontella sinensis* (NC 001713), the three red algae *Cyanidium caldarium* (NC 001840), *Porphyra purpurea* (NC 000925), and *Guillardia theta* (NC 002752), the cyanelle of *Cyanophora paradoxa* (NC 001675), and the homologous genes of the blue–green alga *Synechocystis PCC6803* (NC 000911) (see Table 1 for abbreviations.) The genes used, with the numbers of amino acid sites in the alignment in parenthesis, are *atpA* (504), *atpB* (501), *atpE* (171), *atpF* (191), *atpH* (82), *ccsA* (350), *petA* (347), *petB* (234), *petD* (180), *psaA* (754), *psaB* (738). *psaC* (82), *psaJ* (45), *psbA* (360), *psbB* (509), *psbC* (487), *psbD* (353), *psbE* (84), *psbF* (44), *psbH* (82), *psbI* (54), *psbJ* (42), *psbK* (98), *psbL* (39), *psbN* (44), *psbT* (35), *rbcL* (497), *rpl2* (296), *rpl14* (124), *rpl16* (143), *rpl20* (128), *rpl36* (48), *rpoA* (529), *rpoB* (1654), *rpoC1* (1091), *rpoC2* (1929), *rps2* (279), *rps3* (239), *rps4* (212), *rps7* (154), *rps8* (140), *rps11* (147), *rps12* (126), *rps14* (103), *rps18* (173), *rps19* (94), *ycf4* (195), and *ycf9* (110).

Alignments were performed with Clustal W (Thompson et al. 1994) and written into MOLPHY and PHYLIP format with scripts bundled with MOLPHY (Adachi and Hasegawa 1996). Only the four RNA polymerase genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*) and *ccsA* showed alignments with many gaps.

The number of trees increases very rapidly with the number of taxa. We therefore fixed the uncontentious phylogenetic relationship between the green algae and the land plants to ((nep,chl),(mar,(pin,((ory,zea),(nic,ara))))) and refer to this group as the ''greens.'' That left seven taxa in the analysis (syn, cyp, cyc, gui, por, odo, and the greens), resulting in 945 unrooted trees. To speed up some of the analyses, we reduced this number further by positing that the red algal lineage (gui,por,odo,cyn) is monophyletic; that left 45 unrooted trees.

Log-likelihoods were calculated with two programs, for all 945 trees with ''protml'' (Adachi and Hasegawa 1996) and for the reduced set of 45 trees with ''aaml'' (Yang 1997). For both programs we used the JTT-F model (Jones et al. 1992), where amino acid frequencies are adjusted to those under analysis. As demonstrated previously (Adachi et al. 2000), evolutionary rates of amino acid sites are variable within the 48 genes. In the program ''protml,'' no adjustment is made for variability of evolutionary rates of sites within a gene. In the program ''aaml,'' we selected a model with a Γ-distribution of site rates with eight categories. Considering speed,

we performed the aaml analyses on the subset of 45 trees. Unfortunately, we could obtain RELL bootstrap proportions with aaml only for the shorter genes.

To eliminate part of the site heterogeneity, we processed the data in the following way: we used the Waddell et al. (1999) capture–recapture method to eliminate an excess of invariant sites using the land plants as the monophyletic group. This procedure removed 22% of all sites. Subsequently, sites that varied or required sequence gaps within the land plants, a monophyletic subset of the data with a relatively low proportion of total sequence variation, were eliminated. By this procedure all six land plants were reduced to a single taxon and only 5404 slowly evolving sites remained in the data set. For comparison, we also performed many analyses on data sets with all variable sites (but after removal of the excess invariant sites). The results of the maximum-likelihood single-locus analyses with protml (or aaml) were matrices $\{l_{g,t}\}$, with $1 \le g \le 48$ and $1 \le t \le 945$ (or $1 \le t \le 45$, respectively).

As discussed above, paralogy is quite likely for *rbcL*. We performed many analyses with and without this gene. We note that other genes may also show paralogy but have not been analyzed in sufficient detail or are too short to be informatively analyzed. Furthermore, the probabilistic model below should take paralogy into account.

Phylogenetic inference may be compromised by bias in amino acid composition. Since we concentrate mainly on differences between genes, we use the following metric, denoted r_f, to rank the genes according to their conformity with the average: for a concatenation of all genes as well as for each gene separately, we calculated the **F** matrices (Lockhart et al. 1994) for each pair of taxa. The **F** matrices were normalized to sum to 1. For each gene, sums of the products of each entry in the gene's **F** matrix with that of the concatenated **F** matrix were calculated and summed for all pairwise comparisons of taxa. Genes that are similar to the average will have a high value of the metric.

Consensus trees were calculated using five methods: (i) all genes were concatenated, the maximum log-likelihoods of all 945 trees were evaluated, and RELL bootstrap analysis (Adachi and Hasegawa 1996) was performed with protml; (ii) the same procedure as in i was used but only 45 trees were analyzed with aaml using the discrete Γ option; (iii) the same procedure as in i was used but *rbcL* was left out of the analysis; (iv) the best 20 trees in analysis i were reanalyzed with parsimony criteria using the program "protpars" in PHYLIP (Felsenstein 1993); and (v) the maximum log-likelihoods were calculated independently for all 48 genes and summed subsequently.

A Probabilistic Model for Estimation of the Consensus Tree. To examine the heterogeneity among trees, and the tendency of individual gene trees to conform to the consensus, we employed a probabilistic model that aims to obtain the posterior probability of a tree being the consensus tree and the probability of each gene to follow the consensus tree or its own unique tree. The model is based on three conditional probabilities. The genes are divided into two sets, those that follow the current consensus tree and those that do not. Conditional on the number of genes that follow the consensus tree, we calculate the probability of an arbitrary gene to follow the consensus tree prior to observing its data. Conditional on the data, the current consensus tree, and the prior probability of genes to follow it, we calculate a new phylogeny for all the genes. Conditional on the set of genes that share the phylogenetic history, we calculate a new consensus tree. Cycling through these three steps gives the posterior distribution of the consensus tree and the probability of genes to follow it or to have a different gene tree (key mathematical symbols are listed in Table 2).

Let Y_g be the aligned data set for the gth gene. Let the 945 trees be indexed by t. For our calculations we need the posterior probability of each tree t given the information from each gene Y_g, i.e., $\Pr(t|Y_g)$; this is proportional to the likelihood $\Pr(Y_g|t)$. These probabilities are difficult to estimate and we use two approximations, which are both valid for large samples N (in our case for long sequences): the maximum likelihood (ML) $L_{g,t} = \exp(l_{g,t})$ is approximately proportional to the likelihood, i.e., $\Pr(Y_g|t) \approx const \cdot L_{g,t}$, where *const* is a constant; and the bootstrap proportion (BP) of tree t is approximately equal to the tree's posterior probability: $BP_{g,t} \approx \Pr(t|Y_g)$.

Table 2. Key mathematical symbols

Symbol	Explanation
g	Index of the gene
Y_g	Data of the gth gene
t	Index of the tree
$l_{g,t}$	Maximum log-likelihood of the gth gene and tth tree
$L_{g,t}$	Maximum likelihood of the gth gene and tth tree ($\exp(l_{g,t})$)
BP	Bootstrap proportion (estimated with RELL bootstrap)
c	Current consensus tree
$\boldsymbol{\gamma}$	A characteristic vector that has an entry of 1 for the current consensus tree c and entries 0 for all other trees
$\boldsymbol{\tau} = \{\tau_{g,t}\}$	A characteristic matrix that has entries of 1 if the gth gene follows the tth tree and $\tau_{g,t} = 0$ otherwise
π	Probability of an arbitrary gene to follow the consensus tree prior to observing its data
z_g	An indicator variable; has an entry of 1 if the gth gene is currently following the consensus tree and 0 otherwise

For our data, the ranking of trees provided by the two estimates was very similar, but the ML estimator showed bigger differences in probability between trees compared to the BP estimator. This trend seems to hold also for other data (Table 2 of Shimodeira and Hasegawa [1999]). Since our goal is to detect differences among genes, the BP estimator is more conservative and thus probably more appropriate than the likelihood estimator.

Let the variable c be the current consensus tree and let the characteristic vector $\boldsymbol{\gamma} = \{\gamma_t\}$, with $1 \le t \le 945$, be defined such that $\gamma_t = 1$ if $t = c$ and $\gamma_t = 0$ otherwise. Some genes follow the consensus tree, while others have their own associated phylogenetic tree. Let the characteristic matrix $\boldsymbol{\tau} = \{\tau_{g,t}\}$, with $1 \le g \le 48$ and $1 \le t \le 945$, be defined such that $\tau_{g,t} = 1$ if the gth gene follows the tth tree and $\tau_{g,t} = 0$ otherwise. Partition the genes into two sets, those for which $\tau_{g,c} = 1$ and those for which $\tau_{g,c} = 0$, i.e., the set of genes that follows the consensus tree and the set that does not. Furthermore, let π be the probability that an arbitrary gene will follow the consensus tree prior to having observed its data.

The probabilistic model is specified by three conditional distributions for the vector $\boldsymbol{\gamma}$, the matrix $\boldsymbol{\tau}$, and π. The three conditional distributions uniquely specify the joint posterior distribution of the three variables, if such a distribution exists (Casella and George 1992).

The first conditional distribution is that of $\boldsymbol{\gamma}$, a multinomial generalization of the Bernoulli distribution. Let z_g be an indicator variable that takes the value of 1 if the gth gene is currently following the consensus tree and 0 otherwise. With uninformative priors, the "probability" of tree t (denoted p_t) being selected as the next consensus tree is determined from the product of the likelihoods for all genes with $z_g = 1$ as follows:

$$p_t = \frac{\Pi_g \Pr(t|Y_g)^{z_g}}{\Sigma_t \Pi_g \Pr(t|Y_g)^{z_g}} \quad (1)$$

This is approximated either by $\Pi_g L_{g,t}^{z_g}/\Sigma_t \Pi_g L_{g,t}^{z_g}$ or by $\Pi_g BP_{g,t}^{z_g}/\Sigma_t \Pi_g BP_{g,t}^{z_g}$. Note that selecting a new vector of γ is equivalent to selecting a new consensus tree c.

The second conditional distribution is that of the probability that an arbitrary gene will follow the consensus tree π. This conditional distribution depends only on the current number of genes that follow the consensus tree. This number is binomially distributed, such that the natural distribution for π is the conjugate distribution of the binomial, i.e., the β distribution. With a β prior with $\alpha = \beta = 1$ (Gelman et al. 1995), the conditional distribution of π is

$$\Pr(\pi|z_1, \ldots, z_{48}) \propto \pi^{(\Sigma_g z_g)}(1-\pi)^{(48-\Sigma_g z_g)} \qquad (2)$$

The third conditional distribution is that of the matrix $\tau_{g,t}$. For each gene g, the distribution of $\tau_{g,t}$ depends on the consensus tree c and π and the data Y_g, but is conditionally independent of the other genes and their data. Let t again index the 945 trees and let c be the consensus tree. Before observing its data, a gene will follow the consensus tree with probability π, while with probability $(1-\pi)$ 1 of the other 944 trees will be chosen, *i.e.,*

$$\Pr(t|\pi, c) = \begin{cases} \pi & \text{if} \quad t = c \\ \frac{1}{944}(1-\pi) & \text{if} \quad t \neq c \end{cases}$$

The functional form of the distribution of each gene $\tau_{g,t}$ is a Bernoulli distribution as in the case of γ_t. Using Bayes's rule to combine the "prior" probability $\Pr(t|\pi, c)$ with the likelihood, we obtain the "probability" of tree t (denoted $p_{g,t}$):

$$\begin{aligned} p_{g,t} &= \Pr(Y_g|t)\Pr(t|\pi, c)/\Sigma_t \Pr(Y_g|t)\Pr(t|\pi, c) \\ &\approx L_{g,t}\Pr(t|\pi, c)/\Sigma_t L_{g,t}\Pr(t|\pi, c) \end{aligned} \qquad (3)$$

From eq. (3), a new set $\tau_{g,t}$ can be obtained for each gene g.

These three conditional distributions are sufficiently simple that sampling from them is possible. Hence, the approximate posterior distribution may be obtained by cyclically and iteratively sampling from the conditional distributions, i.e., by a Gibbs sampling scheme (e.g., Casella and George 1992; Gelman et al. 1995).

In our implementation of the Gibbs sampler, the initial conditions are specified as follows: π is sampled with equal probability in the unit interval. Then the genes are partitioned into two sets, the set that follows the consensus tree and the set that does not, by flipping a biased coin with probability π. Conditional on this partition, an initial value for γ, i.e., a consensus tree, is drawn. Then the $\tau_{g,t}$ are sampled conditional on all γ and π. These initial conditions are "overdispersed," i.e., they are more variable than the posterior distribution (Gelman et al. 1995).

The speed of convergence of such overdispersed initial conditions to the posterior distribution allows inference of efficiency of the sampler, such that appropriate numbers of iterations can be chosen. In our case, it was sufficient to discard the first 100 iterations (burn-in), while the next 1000 were treated as samples from the posterior distribution; this process was repeated 50 times.

Check of the Method. To check our methods, we also performed the above analysis on just the green algae and land plants. For this reduced data set, phylogeny is known in advance and paralogy due to differential incorporation into the inverted repeat region does not compromise the results because of the proofreading mechanism that homogenizes the two repeats. Furthermore, taxon sampling is more even than in the whole data set. Hence, we expect only the covarion structure to lead to deviations from the common phylogeny. To reduce the data set to a manageable number of taxa, we left out *Nicotiana* and *Zea*, as another dicot or monocot, respectively, is included in the data set. In this case, we did not eliminate the excessive invariable sites nor did we strip the excessively variable sites.

Table 3. Top five rooted chloroplast trees from concatenated genes—bootstrap proportions using RELL and log-likelihoods (11): (a, b) without Γ correction for site rate heterogeneity; (c, d) with Γ correction for site rate heterogeneity; (a, c) without stripping of hypervariable sites; (b, d) with stripping of hypervariable sites

TREE	RELL	II
	a	
((cyp,greens),((gui,por),(odo,cyc)))	0.775	0.0
((cyp,((gui,por),(odo,cyc))),greens)	0.076	−30.3
((cyp,greens),((gui,(odo,cyc)),por))	0.064	−33.3
(cyp,(greens,((gui,por),(odo,cyc))))	0.050	−33.1
	b	
((cyp,greens),(((gui,odo),por),cyc))	0.019	−70.4
((cyp,greens),((gui,por),(odo,cyc)))	0.357	0.0
((cyp,greens),((gui,odo),(cyc,por)))	0.294	−4.3
((cyp,((gui,por),(odo,cyc))),greens)	0.164	−11.4
((cyp,((gui,odo),(cyc,por))),greens)	0.081	−19.7
((cyp,greens),(((gui,odo),por),cyc))	0.032	−24.2
	c	
((cyp,greens),((gui,por),(odo,cyc)))	0.272	0.0
((cyp,((gui,por),(odo,cyc))),greens)	0.142	−5.6
(cyp,(greens,((gui,por),(odo,cyc))))	0.139	−5.5
((cyp,greens),((gui,odo),(cyc,por)))	0.118	−13.4
(cyp,(greens,((gui,odo),(cyc,por))))	0.053	−19.9
	d	
((cyp,((gui,por),(odo,cyc))),greens)	0.191	0.0
((cyp,greens),((gui,odo),(cyc,por)))	0.187	−0.5
((cyp,((gui,odo),(cyc,por))),greens)	0.182	−0.3
((cyp,greens),((gui,por),(odo,cyc)))	0.151	−1.2
(cyp,(greens,((gui,por),(odo,cyc))))	0.044	−8.8

Results

In the analysis of the concatenation of all genes (Table 3), we recovered many phylogenetic trees by RELL bootstrap analysis. All trees, however, support monophyly of the red/brown algal group (gui, por, cyc, odo; short reds). Relationships within the reds are variable, as is the position where the outgroup syn joins the three major ingroups (greens, reds, and cyp), i.e., the position of the root. Estimated branch lengths for the contentious branches are short (e.g., Fig. 1a); the exterior branches (leading to leaves) among the reds are about 10 times longer than the interior ones and have much smaller standard errors, with respect to length.

Three of the four analyses support the same most probable tree (Fig. 1a), which we refer to as Tree 1; the support is strongest in the analysis without a Γ distribution and without stripping of excessively variable sites (Table 3a). This tree unites the cyanelle with the green algal group—a relationship that corresponds to Tree 3 of Adachi et al. (2000)—while there are two sister groups within the reds: por–gui and cyc–odo. Stripping of excessively variable sites reduces the amount of information in the data. Probably for this reason, differences in likelihood and

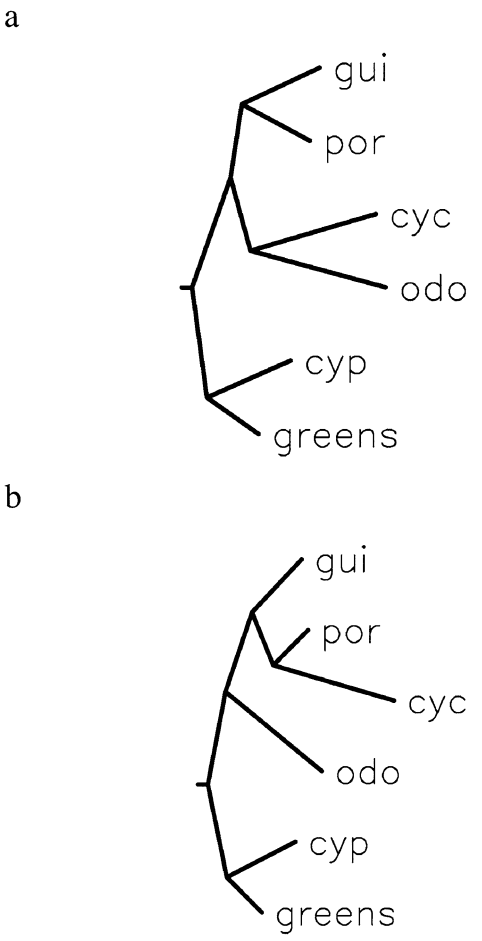

Fig. 1. (**a**) The most probable consensus tree and (**b**) the *atpA*, *atpF*, and *atpH* tree; both trees rooted with *Synechocystis* as the outgroup. The horizontal extent of the branches corresponds to the branch length.

bootstrap proportions among trees are less pronounced with site-stripping (Table 3b), yet the general results are similar. The same relationship within the reds is also favored by the analysis with Γ correction of variability among sites but without site-stripping (Table 3c). With this analysis, however, support for the sister-group relationship of cyp and greens is waning; in fact, none of the three possibilities of how syn joins the ingroup is convincingly excluded. Discrimination between trees is even more difficult in the analyses with both Γ correction and site-stripping (Table 3d): neither relationships within the reds nor the position of the root is resolved. While likelihood and RELL bootstrap proportions agree on the relative ranking of trees, using the likelihood ratio criterion to discriminate between trees suggests much more confidence in the best tree than the RELL bootstrap proportions.

Tree 1 was also the best in an analysis without *rbcL*, with and without site-stripping but without the Γ correction (data not shown). Parsimony analyses of stripped and unstripped data, however, favored different trees (data not shown).

Phylogenetic inference may be compromised by differences among genes in evolutionary rates and amino acid bias. The former problem can be tackled with separate analyses of genes and summation of the likelihoods. Results from this procedure are similar to that of the concatenation (data not shown). But from this analysis, it was evident that some genes preferred very different trees over the consensus tree: e.g., with *atpA*, *petA*, and *petB*, the bootstrap proportion of all trees with a sister-group relationship of por–gui and cyc–odo within the reds combined was way below 0.01. In fact, data from *atpA* suggest a different phylogeny within the reds (Fig. 1b). The highly variable RNA polymerases and *petB* favored paraphyletic or polyphyletic grouping of the reds (data not shown), despite the fact that their extraordinary length should have provided much information. On the other hand, *petD* and *psaB* favored a sister-group relationship of por–gui and cyc–odo within the reds.

To address the problem of differential amino acid bias, we use a metric to quantify the concordance of amino acid bias of individual genes with that of the concatenated sequence. Without stripping of excessively variable sites, the RNA polymerases, *atpE*, *atpF*, *ccsA*, *rps11*, *rps19*, and *ycf4* are most different from the concatenation, whereas *petB*, *petD*, and *psbA* are most similar to the concatenation (Table 4). With site-stripping, *atpF*, *ccsA*, *petA*, *rpl36*, *rps11*, *ycf4*, and *ycf9* and, only to a lesser degree, the RNA polymerases are most different from the concatenation, while *psbH*, *psbK*, and *rps12* are most similar to the concatenation (Table 4).

With the probabilistic model and Gibbs sampling, the most selected tree was again the same as the one selected by concatenation of genes without taking into account variable rates with the Γ model (Table 5), although the sister-group relationship between cyp and the greens is less strongly supported. In accordance with the concatenated analysis, no trees were recovered that contradicted monophyly of the reds. With the Γ correction and analyzing only 45 trees, different trees are selected: with the full data set, relationships within the reds are resolved differently, while with site-stripping, relationships among the major groups ore resolved differently.

The posterior probability that a gene will follow a certain consensus tree was also evaluated with the probabilistic model. Conditional on a certain tree being selected, all methods of data processing (with and without site-stripping and with and without Γ correction) give consistent results. When using the ML estimator of the posterior probability of a tree given the data, the number of genes following a tree is surprisingly low—less than half of the genes support any single tree (Tables 6a, b, e, f). When using the RELL bootstrap proportion estimator, the number of genes following the consensus tree is higher. Only a few

Table 4. Index for conformity with consensus amino acid bias, r_f (a higher index indicates higher conformity): first column, full data set; second column, excessively variable sites stripped

Gene	r_f	red.
atpA	2.01	0.93
atpB	1.85	0.87
atpE	1.55	0.78
atpF	1.21	0.51
atpH	2.38	0.94
ccsA	1.31	0.66
petA	1.63	0.66
petB	2.45	1.02
petD	2.59	1.10
psaA	2.29	0.97
psaB	2.16	0.94
psaC	1.95	0.83
psaJ	1.92	0.89
psbA	2.49	1.03
psbB	2.18	0.95
psbC	2.37	0.99
psbD	2.39	1.05
psbE	1.82	0.73
psbF	2.24	0.92
psbH	2.00	1.14
psbI	2.44	1.08
psbJ	2.09	0.78
psbK	2.16	1.18
psbL	1.85	0.80
psbN	1.91	0.81
psbT	2.42	1.02
rbcL	1.19	0.77
rpl2	2.00	0.96
rpl14	1.93	0.95
rpl16	1.53	0.85
rpl20	1.64	0.75
rpl36	1.74	0.62
rpoA	1.49	0.77
rpoB	1.55	0.75
rpoC1	1.43	0.69
rpoC2	1.16	0.67
rps2	1.63	0.86
rps3	2.33	1.11
rps4	1.59	0.71
rps7	1.71	0.87
rps8	1.66	0.80
rps11	1.50	0.69
rps12	1.66	0.92
rps14	1.57	0.72
rps18	1.69	0.70
rps19	1.53	0.72
ycf4	1.42	0.60
ycf9	1.72	0.57

Table 5. Top five rooted chloroplast trees selected by the Gibbs sampling as scheme—posterior probability (p): (a–d) without Γ correction for site rate heterogeneity; (e, f) with Γ correction for site rate heterogeneity; (a, c, e) without stripping of hypervariable sites; (b, d and f) with stripping of hypervariable sites

Tree	
a	
((cyp,greens),((gui,por),(odo,cyc)))	0.435
(cyp,(greens,((gui,por),(odo,cyc))))	0.126
((cyp,((gui,odo),(cyc,por))),greens)	0.089
((cyp,((gui,por),(odo,cyc))),greens)	0.085
(cyp,(greens,(((gui,odo),por),cyc)))	0.051
b	
((cyp,greens),((gui,por),(odo,cyc)))	0.240
((cyp,greens),(((gui,odo),por),cyc))	0.145
((cyp,greens),((gui,odo),(cyc,por)))	0.112
((cyp,((gui,por),(odo,cyc))),greens)	0.098
((cyp,(((gui,odo),por),cyn)),greens)	0.059
c	
((cyp,((gui,por),(odo,cyc))),greens)	0.325
((cyp,greens),((gui,por),(odo,cyc)))	0.197
((cyp,((gui,odo),(cyc,por))),greens)	0.099
(cyp,(greens,(((gui,odo),por),cyc)))	0.065
(cyp,(greens,((gui,por),(odo,cyc))))	0.062
d	
((cyp,((gui,por),(odo,cyc))),greens)	0.537
((cyp,((gui,odo),(cyc,por))),greens)	0.070
(cyp,(greens,(((gui,odo),por),cyc)))	0.068
(cyp,(greens,((gui,por),(odo,cyc))))	0.051
(cyp,(greens,((gui,odo),(cyc,por))))	0.070
e	
((cyp,greens),(((gui,por),odo),cyn))	0.318
((cyp,((gui,odo),(cyn,por))),greens)	0.077
(cyp,(greens,((gui,odo),(cyn,por))))	0.075
(cyp,(greens,((gui,por),(odo,cyn))))	0.068
((cyp,greens),((gui,por),(odo,cyc)))	0.062
f	
((cyp,((gui,odo),(cyc,por))),greens)	0.182
((cyp,greens),(((gui,por),odo),cyc))	0.089
((cyp,greens),((gui,odo),(cyc,por)))	0.080
((cyp,greens),((gui,por),(odo,cyc)))	0.068
(cyp,(((gui,odo),(cyn,por)),greens))	0.062

genes, e.g., *petA*, *petB*, and *rpoB*, do not follow the consensus tree in the majority of cases. A higher number of genes following the consensus seems biologically more plausible than the results with the likelihood estimator. There seems to be no correlation of the posterior probability of a gene to follow the most selected tree and the gene's deviation in amino acid bias from the consensus, e.g., *ycf4* and *rpoA* show a bias deviating from the concatenation and the former follows the consensus tree while the latter does not. The genes *petB* and *petD*, which have an amino acid bias similar to the concatenation and are strictly syntenic and functionally similar, have very different preference with respect to the consensus tree. The only trend that we are able to discern is that genes that have a strongly deviating bias are more sensitive to data processing and method of analysis.

The case of *petB* and *petD*, which are syntenic and similarly functionally constrained yet follow very different trees, is no exception. We cannot detect a tendency for syntenic genes and/or similarly functionally constrained genes to follow the same trees with one possible exception: *atpA*, *atpF*, and *atpH* seem to follow the same tree (Fig. 1b), although *atpF* also follows the most probable consensus tree. This

Table 6. Posterior probability of genes to follow the most probable consensus tree, with and without stripping of hypervariable sites (+ss and −ss, respectively) and with and without the Γ model for varying evolutionary rates (+Γ and −Γ, respectively), using RELL bootstrap proportions (*R*) or maximum likelihood (*L*) to calculate the posterior probability of trees

Gene	−ss−Γ *l*	+ss− Γ *l*	−ss−Γ *r*	+ss−Γ *r*	−ss+Γ *l*	+ss+Γ *l*
atpA	0.00	0.16	0.42	0.53	0.00	0.00
atpB	0.00	0.00	0.81	0.90	0.01	0.27
atpE	0.13	0.78	0.86	0.89	0.35	0.55
atpF	0.96	0.41	0.97	0.98	0.34	0.29
atpH	0.00	0.00	0.32	0.29	0.00	0.00
ccsA	0.04	0.40	0.25	0.22	0.18	0.45
petA	0.00	0.00	0.39	0.35	0.00	0.00
petB	0.00	0.00	0.26	0.23	0.00	0.00
petD	0.99	0.99	0.99	0.99	0.90	0.85
psaA	0.00	0.00	0.33	0.63	0.00	0.00
psaB	0.68	0.98	0.51	0.78	0.49	0.88
psaC	0.74	0.84	0.28	0.28	0.71	0.85
psaJ	0.00	0.38	0.24	0.22	0.04	0.56
psbA	0.88	0.87	0.95	0.96	0.97	0.94
psbB	0.00	0.02	0.95	0.94	0.00	0.00
psbC	0.06	0.11	0.81	0.76	0.06	0.07
psbD	0.51	0.96	0.56	0.69	0.23	0.81
psbE	0.25	0.49	0.44	0.58	0.16	0.41
psbF	0.63	0.92	0.41	0.71	0.38	0.65
psbH	0.98	0.00	0.90	0.93	0.40	0.12
psbI	0.94	0.84	0.32	0.30	0.44	0.37
psbJ	0.78	0.38	0.96	0.96	0.14	0.03
psbK	0.39	0.22	0.64	0.71	0.23	0.23
psbL	0.10	0.02	0.30	0.28	0.04	0.03
psbN	0.85	0.65	0.38	0.54	0.55	0.45
psbT	0.41	0.43	0.26	0.25	0.20	0.19
rbcL	0.67	0.90	0.96	0.98	0.51	0.58
rpl12	0.75	0.30	0.68	0.71	0.27	0.21
rpl14	0.97	0.58	0.99	0.99	0.40	0.27
rpll6	0.95	0.97	0.88	0.82	0.25	0.61
rpl20	1.00	0.89	0.98	0.98	0.85	0.79
rpl36	0.39	0.36	0.32	0.29	0.47	0.47
rpoA	0.04	0.00	0.97	0.98	0.16	0.25
rpoB	0.00	0.00	0.33	0.29	0.25	0.27
rpoCl	0.83	0.98	0.98	0.98	0.47	0.66
rpoC2	0.00	0.76	0.90	0.89	0.66	0.55
rps2	0.65	0.72	0.93	0.95	0.12	0.27
rps3	0.99	0.99	0.95	0.95	0.82	0.79
rps4	0.65	0.88	0.76	0.83	0.27	0.36
rps7	0.95	0.28	0.96	0.96	0.60	0.47
rps8	0.88	0.08	0.78	0.82	0.62	0.09
rps11	0.35	0.08	0.31	0.28	0.19	0.03
rps12	0.02	0.01	0.25	0.37	0.02	0.05
rps14	0.04	0.06	0.95	0.94	0.07	0.11
rps18	0.99	0.94	0.95	0.96	0.86	0.82
rps19	0.62	0.79	0.91	0.81	0.56	0.70
ycf4	0.88	0.41	0.99	0.99	0.36	0.05
ycf9	0.99	0.96	0.96	0.96	0.59	0.49

observation should, however, be confirmed by analysis of other genes in the same syntenic group that are left out of this analysis because they are lost or transferred to the nucleus in the green algae and land plants.

Check of the Method. The results of the analysis of only the green algae and land plants contrast very strongly with those from the full data set: the known true phylogeny, ((nep,chl),mar,(pin,(ory,ara))), was recovered as the consensus tree in more than 95% of the cases. Furthermore, with RELL-bootstrap proportions, all genes supported the consensus tree in more than 50% of the cases (data not shown). With posterior probabilities estimated from log likelihoods, recovery of the consensus was again above 95%, and the support of the consensus tree was higher than 50% for all but two genes.

Discussion

Gene histories may be divergent within a genome. Our goal was to identify congruence or incongruence among gene trees using the plastid genome as an example. Specifically, we tried to identify subsets of genes exhibiting the same phylogeny, as opposed to others exhibiting varying phylogenies.

Information from whole plastid genomes of 14 taxa was used to reconstruct gene trees and a consensus tree and to compare the gene trees. The 14 taxa are 6 land plants, 2 green algae, a diatom, 2 rhodophytes, a cryptophyte, the cyanelle of the glaucocystophyte *Cyanophora*, and the blue–green alga *Synechocystis* as an outgroup. We used the noncontentious relationships among the green algae and the six land plants for data processing (site-stripping of excessively variable sites using the method of Waddell et al. [1999]) and to reduce the data set to seven taxa. We then inferred the likelihood of 945 trees for each of the 48 protein coding genes using maximum likelihood methods and bootstrap proportions.

Is there a Consensus Tree?

From earlier analyses with single genes or small groups of genes and from anatomical and biochemical evidence, we would expect *a priori* that the green algae and land plants form a monophyletic group and that the cyanelle of *Cyanophora* is quite isolated (e.g., Martin et al. 1998; Glöckner et al. 2000; Douglas and Penny 1999). The question of the monophyly of the two red algae, the cryptophyte, and the diatom is more contentious (Martin et al. 1998; Lockhart et al. 1999), but gene arrangements and sequence analysis indicate monophyly (e.g., Martin et al. 1998; Glöckner et al. 2000; Douglas and Penny 1999). On the other hand, the relationships of the three major groups of chloroplasts (the red/brown algae, the green algae/land plants, and the cyanelle *Cyanophora*) to the outgroup (the blue–green alga *Synechocystis*) are contentious. Similarly contentious are the relationships within the red/brown algae, but the similar gene order suggests a close relationship between *Guillardia* and *Porphyra* (Glöckner et al. 2000; Douglas and Penny 1999). The relationship of the major groups was analyzed previously with a slightly smaller data set (Adachi et al. 2000). This study favored a sister-group relationship of the red/brown algae and the green algae/land plants with *Cyanophora* branching off first without being able to exclude the other two possibilities. Hence, our analysis might (i) clarify the relationships of *Odontella*, the red algae, and the cryptophyte, i.e., the relationship within the red/brown algae, and (ii) the relationship among the major groups: the red/brown algae, the green algae, and the cyanelle.

In a RELL bootstrap analysis, we recovered only trees that supported monophyly of the following group: the two red algae, the diatom, and the cryptophyte, i.e., the red/brown algae. But contrary to the earlier analyses (Martin et al. 1998; Adachi et al. 2000), we cannot find a single, dominant consensus tree or a set of closely related trees. The most likely trees differ significantly in the resolution of the relationships within the red/brown algae and of the relationships between the three major chloroplast groups and the outgroup. Even if only the relationships of the major groups are considered, we find less support than Adachi et al. (2000) did for the hypothesis of a sister-group relationship of red algae and green algae/land plants and that *Cyanophora* branches off first. Instead, our analysis favors a sister-group relationship between *Cyanophora* and the green algae/land plants. This difference from Adachi et al. (2000) seems to be caused exclusively by the differences in taxa analyzed: with the same taxa as used by Adachi et al. (2000), we achieve nearly identical results despite slight differences in data processing (data not shown). The problem of the phylogeny of the red/brown algae did not occur in the Adachi et al. (2000) analysis, as the chloroplasts of *Cyanidium caldarium* and *Guillardia theta* were not available for their analysis.

In summary, our attempt to reconstruct a consensus tree from the whole data set provides clear evidence for a monophyletic red/brown algal group but can neither resolve the phylogeny within this red/brown algal group nor clarify the basic relationships of *Cyanophora*, the red/brown algae, and the green algae/land plants with the outgroup *Synechocystis*. Furthermore, analysis of single genes indicates that some genes actually reject the consensus tree (bootstrap proportion $\ll 0.01$) and prefer entirely different trees.

Are the Gene Trees Congruent?

The results from the previous subsection, led us to abandon the hypothesis that most genes have similar phylogenetic signals i.e., follow the consensus tree. Alternative hypotheses include horizontal gene transfer (HGT), gene duplication and loss (paralogy), and differences among genes in amino acid bias and/or covarion structure, i.e., selection. For *rbcL*, ample evidence of HGT or paralogy has been compiled previously (Delwiche and Palmer 1996). Later, Martin and Schnarrenberger (1997) explained the phylogenetic pattern found for *rbcL* by paralogy. In an analysis of the *secA* gene, Valentin (1997) found a similarly deep split between red and brown algae and

land-plant chloroplasts and argued on this basis for a polyphyletic origin of chloroplasts. In a reanalysis of *secA* data, Barbrook et al. (1998) found the support for polyphyly of the chloroplasts decreasing if unequal AT content was accounted for by LogDet transformation of distances (Lockhart et al. 1994). For the RNA polymerases *rpoB*, *rpoC1*, and *rpoC2* and possibly for *rps8*, Martin et al. (1998) found evidence for deviation from the consensus tree. In a reanalysis of the same data, Lockhart et al. (1999) found similarly conflicting phylogenetic signals for *rpoB*, *rpoC1*, and *rpoC2* and the rest of the genes. They questioned the phylogenetic signal provided by the rest of the genes, because taxa that come out closely related in the analysis (*Odontella* and *Porphyra*) share similar amino acid biases. On the other hand, related taxa are also more likely to share amino acid biases: thus the phylogenetic signal may be genuine. Furthermore, as noted by Adachi et al. (2000) and also apparent in the current study, the RNA polymerases are quite deviant from other genes in evolutionary rate and amino acid composition.

Genes that have a different phylogenetic signal hamper estimation of the consensus tree. Excluding these genes from the analysis should therefore improve estimation of the consensus tree. But, with the possible exception of *rbcL*, we do not know *a priori* which genes follow the consensus phylogeny and which exhibit their own phylogenetic signal. Therefore, we formulate a probabilistic model where we split the genes into two sets, those that follow the consensus phylogeny and those that have their own phylogeny. Membership in these two classes, the consensus tree, and all phylogenetic trees associated with independent genes are estimated concurrently via a Gibbs sampling scheme. Because many consensus trees may be imputed, subsets of genes that follow the different trees are selected.

In this analysis, the most supported tree is the tree with the highest RELL bootstrap proportion for the concatenation of all genes. Data processing and method of analysis only weakly influence the posterior probability of a particular gene to follow a particular tree. Rather, the subset of genes selected conditional on a particular consensus tree is quite stable.

There is little evidence that genes in conserved collinear groups or functionally related genes have similar phylogenetic patterns. For example, *petB* and *petD* are functionally related, in the same collinear group, and have similar amino acid biases, yet follow different trees. But possibly the collinear group containing *atpA*, *atpF*, and *atpH* follows the same tree (Fig. 1b), although *atpF* and *atpH* are too short and uninformative to provide a clear answer at the present level of analysis. Exclusion of more distantly related taxa (green algae and land plants) would increase the number of genes in that short collinear stretch and thus increase resolution.

This pattern of incongruence of gene trees with the whole data set is in striking contrast to the pattern of congruence when the data set is reduced to the monophyletic group of green algae and land plants (the greens). Within the greens, better taxon sampling, closer phylogenetic distances among taxa, and absence of differentiation among the inverted repeats may produce these contrasting, benign results. Which of these factors or which combination of factors is responsible for the concordance of gene trees in the greens and the discordance in the whole data set is open.

How to Explain the Incongruence of Gene Trees?

In an attempt to explain the observed pattern, we concur with Martin and Schnarrenberger (1997) that HGT is unlikely. Furthermore, amino acid bias only weakly influences the phylogenetic pattern. Another possibility is that genes may differ in their covarion structure (Lockhart et al. 1998). This is rather difficult to model. Furthermore, with a strong covarion structure, gene trees within the green algae and land plants should be similarly discordant as those in the whole data set. The existence of the two inverted or direct repeat units (IRs/DRs) and the fact that copyediting is imperfect in all but the land plants and green algae offer a mechanism for repeated gene duplication via inclusion of new genes into the IRs or DRs and subsequent loss or silencing of one or the other copy. As with all processes amenable to phylogenetic analysis, these events would have to occur at just the right frequency: too high a frequency would swamp the signal with noise, while too low a frequency would lead to too few detectable events. From the stability of gene order and orientation in the IRs/DRs over long evolutionary times (Turmel et al. 1999), variability may be within the right order of magnitude.

Adequate analysis of the possibility of paralogy in the IRs/DRs requires concurrent analysis of information on gene arrangements and amino acid sequence. Gene arrangements may also provide further evidence of phylogenetic relationships. But the most promising direction toward clarification of chloroplast relationships is to increase the data quality, either by increasing the number of genes by excluding the green algae and land plants that have lost many genes shared among the other taxa or by obtaining sequence information on new taxa that help to break up some of the long branches. It is unlikely, however, that the addition of new taxa will overthrow the major result of this paper: many chloroplast genes do not follow the same phylogenetic patterns as the majority of genes.

Widespread discordance of gene trees has.been demonstrated for the lower splits of the tree of life and explained by HGT (e.g., Doolittle 1999; Woese 2002). In precellular evolution before the "Darwinian threshold" (Woese 1998), HGT is very likely. Chloroplast evolution, however, seems to have started after the establishment of cells. Therefore, we think it is unlikely that HGT causes the observed pattern. Instead paralogy is a much more likely candidate mechanism for the observed phenomena.

Acknowledgments. We thank our colleagues at UCR, especially Peter Morrell for discussion and Peter Morrell, Thomas Städler, and Tina Hambuch for critical reading of the manuscript. Comments of two referees helped to improve the manuscript. This work was supported in part by a UCR startup grant and NSF Grants CCR-9988353 and ITR-0085910.

References

Adachi J, Hasegawa M (1996) Computer science monographs, No. 28. MOLPHY Version 2.3: Programs for Molecular Phylogenetics based on Maximum Likelihood. Institute of Statistical Mathematics, Tokyo

Adachi J, Waddell P, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348–358

Barbrook AC, Lockhart JC, Howe CJ (1998) Phylogenetic analysis of plastid orgins based on it secA sequences. Curr Genet 34:336–341

Butterfield N (2000) *Bangiomorpha pubescens* n. gen., n. sp.: Implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology 26:386–404

Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46:167–174

Delwiche C, Palmer J (1996) Rampant horizontal transfer and duplication of rubisco genes in Eubacteria and plastids. Mol Biol Evol 13:873–882

Doolittle W (1999) Lateral genomics. Trends Cell Biol 9:M5–M8

Douglas S, Penny S (1999) The plastid genome of the cryptophyte alga, *Guillardia theta:* Complete sequence and conserved synteny groups confirm its common ancestry with Red Algae. J Mol Evol 48:2367–244

Eisen J (2000) Horizontal gene transfer among microbial geneomes: New insights form complete genome analysis. Curr Opin Genet Devel 10:606–611

Felsenstein J (1993) PHYLIP (phylogeny inference package), manual version 3.5c. Department of Genetics, University of Washington, Seattle

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman and Hall, London

Glöckner G, Rosenthal A, Valentin K (2000) The structure amd gene repertoire of an ancient red algal plastid genome. J Mol Evol 51:382–390

Jones D, Taylor W, Thornton J (1992) The rapid generation of mutation data matrices from protein sequences. J Mol Evol 31:151–160

Knoll A (2000) The early evolution of eukaryotes: A geological perspective. Science 256:622–627

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ (1998) A covariotide model explains apparent phylogenetic structure of photosynthetic lineages. Mol Biol Evol 15:1183–1188

Lockhart PJ, J HC, Barbrook AC, Larkum WD, Penny D (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol 16:573–576

Martin W, Schnarrenberger C (1997) The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: A case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet 32:1–18

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik K (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165

Melkonian M (1996) Systematics and evolution of the algae: Endocytobiosis and evolution of major algal lineages. In: Progress in Botany, Vol 57. Springer-Verlag, Berlin, pp 281–311

Morton B (1998) Selection and the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449–459

Ochman H, Moran N (2001) Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. Science 292:1096–1098

Rosenberg M, Kumar S (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci USA 98:10751–10756

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

Thompson J, Higgins D, Gibson T (1994) Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. Nucleic Acids Res 22:4673–4680

Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea:* Insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci USA 96:10248–10253

Valentin K (1997) Phylogeney and expression of the *secA* gene from a chromophytic alga—Implications for the evolution of plastids and sec-dependent protein translocation. Curr Genet 32:300–307

Waddell P, Cao Y, Hauf J, Hasegawa M (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-log Det plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Syst Biol 48:31–53

Woese CR (1998) On the evolution of cells. Proc Nat Acad Sci USA 95:8742–8747

Woese CR (2002) The universal ancestor. Proc Nat Acad Sci USA 99:8742–8747

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556