J Mol Evol (2000) 50:348–358
DOI: 10.1007/s002399910038

JOURNAL OF MOLECULAR EVOLUTION

Plastid Genome Phylogeny and a Model of Amino Acid Substitution for Proteins Encoded by Chloroplast DNA

Jun Adachi,[1,*] Peter J. Waddell,[2,] William Martin,[3,***] Masami Hasegawa[2]**

[1] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom
[2] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan
[3] Institute für Genetik, Technische Universität Braunschweig, Spielmannstr. 7, D-38023 Braunschweig, Federal Republic of Germany

Received: 3 June 1999 / Accepted: 26 November 1999

Abstract. Maximum likelihood (ML) phylogenies based on 9,957 amino acid (AA) sites of 45 proteins encoded in the plastid genomes of *Cyanophora*, a diatom, a rhodophyte (red algae), a euglenophyte, and five land plants are compared with respect to several properties of the data, including between-site rate variation and aberrant amino acid composition in individual species. Neighbor-joining trees from AA LogDet distances and ML analyses are seen to be congruent when site rate variability was taken into account. Four feasible trees are identified in these analyses, one of which is preferred, and one of which is almost excluded by statistical criteria. A transition probability matrix for the general reversible Markov model of amino acid substitutions is estimated from the data, assuming each of these four trees. In all cases, the tree with diatom and rhodophyte as sister taxa was clearly favored. The new transition matrix based on the best tree, called cpREV, takes into account distinct substitution patterns in plastid-encoded proteins and should be useful in future ML inferences using such data. A second rate matrix, called cpREV*, based on a weighted sum of rate matrices from different trees, is also considered.

* *Present address:* The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan
** *Present address:* Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand
*** *Present address:* Institute of Botany III, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany
Correspondence to: Masami Hasegawa; *e-mail:* hasegawa@ism.ac.jp

Key words: Maximum likelihood — General reversible Markov model — Amino acid substitution — Chloroplast DNA — AA LogDet

Introduction

The maximum likelihood (ML) method is widely used in molecular phylogenetics (Felsenstein 1981). The utility of the method depends on the availabiltiy of realistic models for the change of nucleotide or amino acid sequences during evolution. If one is analyzing DNA sequence data of protein-encoding genes, use of codon-based models of nucleotide substitutions (e.g., Muse and Gaut 1994; Goldman and Yang 1994; Yang et al. 1998) has some advantages over using nucleotide sequences without codon structure. However, if one is interested in estimating very ancient branching events, analyses of amino acid sequences are preferable because synonymous substitutions that are already saturated contain no phylogenetic information and because amino acid substitutions are somewhat easier to model and analyze than codon substitutions. ML analysis of amino acid sequences was first implemented by Kishino et al. (1990) and was further developed with the program ProtML in MOLPHY (Adachi and Hasegawa 1996b).

Transition matrices of amino acid substitutions were first estimated by the parsimony method for data sets that consist mainly of nuclear-encoded proteins; that is, the Dayhoff model (Dayhoff et al. 1978) and the JTT model (Jones et al. 1992). These models are implemented in the ProtML program. It is useful to consider the ML estima-

tion of transition matrices for such data sets and what it can tell us (Reeves 1992). For example, Adachi and Hasegawa (1996a, 1996b) estimated the rate matrix of mitochondrial proteins (mtREV model), which is quite distinct due to a different genetic code to nuclear DNA, and for this reason it has been a popular model to use (e.g., Janke et al. 1997; Zardoya et al. 1998; Cao et al. 1998a, 1998b; Waddell et al. 1999). Estimation of rate matrices via ML is expected to result in less bias than via parsimony (e.g., Perna and Kocher 1995).

Protein-encoding genes of chloroplast DNA (cpDNA) are used widely in plant phylogenetics (e.g., Soltis et al. 1992; Delwiche and Palmer 1997), but amino acid substitution models tuned for such analyses are not yet available. Recently, Martin et al. (1998) estimated the molecular phylogeny of chloroplasts using 45 cpDNA-encoded proteins from nine species encompasing green plants (chlorophytes), a euglenophyte, a red algae (rhodophytes), a diatom, and *Cyanophora* with cyanobacteria as an outgroup. The sequences used provide us with sufficient data for estimating a reliable transition probability matrix of amino acids in chloroplast proteins. In doing so, we need also consider critically the tree relating these sequences.

Many evolutionary studies of plastid phylogeny have been performed with ribosomal RNA (Bhattacharya 1997). The rRNA data has about 1,500 aligned positions and four possible states per position. On the other hand, proteins in chloroplast genomes contain about 10,000 sites for comparison and 20 possible states per position. It seems likely that proteins in chloroplast DNA contain much useful information for studying plastid evolution. Previous work has suggested that this might be the case (Martin et al. 1998), but there have been dissenting views (e.g., Lockhart et al. 1999). Certainly, using this information requires suitable statistical models of the substitution process if the inferred phylogeny is to be reliable. Although it can be computationally demanding when a large number of species are compared, the ML approach has some advantages in phylogenetic inference (e.g., Swofford 1996), and its utility can be steadily extended by the development of more realistic amino acid substitution models (e.g., Adachi and Hasegawa 1996a).

In the absence of well tuned ML substitution models, and the possibility of unequal (nonstationary) amino acid composition in different taxa, it is useful to consider the use of amino acid (AA)–based LogDet distances (Lockhart et al. 1994; Penny et al. 1999; Waddell et al. 1999) combined with a distance based phylogenetic method. If site rate heterogeneity is also expected, then use of the invariant sites–LogDet is also wise (Swofford 1996; Waddell and Steel 1997; Waddell et al. 1999). Similar methods have recently been applied to this data by Lockhart et al. (1999), and we compare and contrast our results with theirs.

Materials and Methods

The data used were complete cpDNA sequences of the land plants *Zea mays* (database accession number: X86563), *Oryza sativa* (X15901), *Nicotiana tabacum* (S54304), *Pinus thunbergii* (D17510), and *Marchantia polymorpha* (X04465); the euglenophyte *Euglena gracilis* (X70810); the rhodophyte *Porphyra purpurea* (U38804); the diatom *Odontella sinensis* (Z67753) and *Cyanophora paradoxa* (U30821); and the complete genome sequence of the cyanobacteria *Synechocystis* PCC6803 as the outgroup. A total of 9,957 amino acids (excluding gapped sites) are used from the following 45 proteins found in all 10 genomes (numbers in parentheses are numbers of amino acid sites); *atpA* (494), *atpB* (473), *atpE* (128), *atpF* (174), *atpH* (81), *petB* (215), *petG* (36), *psaA* (748), *psaB* (726), *psaC* (81), *psaJ* (36), *psbA* (343), *psbB* (507), *psbC* (457), *psbD* (350), *psbE* (72), *psbF* (38), *psbH* (57), *psbI* (36), *psbJ* (38), *psbK* (44), *psbL* (38), *psbN* (43), *psbT* (31), *rpl*14 (120), *rpl*16 (134), *rpl*2 (271), *rpl*20 (112), *rpl*22 (111), *rpl*36 (36), *rpoB* (982), *rpoC*1 (554), *rpoC*2 (762), *rps*11 (126), *rps*12 (123), *rps*14 (98), *rps*18 (58), *rps*19 (91), *rps*2 (223), *rps*3 (202), *rps*4 (193), *rps*7 (155), *rps*8 (130), *ycf*4 (172), and *ycf*9 (58). This is the same data set used in Martin et al. (1998) except for one minor change in the alignment of *petG* (where the number of amino acid sites used is now 36 instead of 37). The total number of sites is used was typically 9,957 (analyses of alternative edits of the data indicate the total number of sites in the text).

In maize, it is estimated that about 25 cpDNA sites are subject to RNA editing (Maier et al. 1995). Although the total degree of editing is not known for all the species in the present data set, it will only affect comparisons of differentially edited codons at the small fraction of edited sites. This minor, and presently difficult-to-gauge effect was also ignored as in Martin et al. (1998) and others using such data.

To estimate the transition probability matrix of the general reversible Markov model of amino acid substitution for cpDNA-encoded proteins, the phylogenetic relationships among the species must be established, or, if there exists any uncertainty, it should be taken into account in estimating the matrix. A previous ML analysis of the same data set used here (Martin et al. 1998) detected only four trees among the 10,000 bootstrap replications with RELL (Kishino et al. 1990; Hasegawa and Kishino 1994), using the JTT-F model (Jones et al. 1992; option "-F" adjusts the amino acid frequencies to those of the data under analysis). Tree-1 with the rhodophyte/diatom grouping gave the highest likelihood. Additional ML analyses are carried out assuming site-heterogeneity, which was not taken into account in Martin et al. (1998). These assumed a Γ-distribution of site rates (Yang 1994; Waddell et al. 1997) and used the AAML program in the PAML package (Yang 1997) with eight categories in the discrete Γ-approximation.

As shown later, pairwise tests of amino acid composition (Penny et al. 1999; Waddell et al. 1999) show clearly nonstationary evolution in these sequences, and this variation is not taken into account in the ML analysis. Therefore, we estimated distances by the LogDet transform (Barry and Hartigan 1987; Lockhart et al. 1994; Lake 1994) applied to amino acid sequences using the program from Waddell et al. (1999), after removal of the invariant sites in proportion to the unvaried sites (Waddell and Steel 1997). The proportion of invariant sites was estimated using the capture-recapture method of Waddell (1995) and Waddell et al. (1999). From these distances, trees were built using the NJ method (Saitou and Nei 1987).

Last leaving rapidly changing sites in any analysis can lead to errors. To help avoid this possibility we used "site stripping" as implemented in Waddell et al. (1999). Such analyses involve removing all amino acid positions showing any change within a defined monophyletic group(s). Such objectively edited data can then be used for any type of analysis (e.g., ML, parsimony, distance-based methods).

The transition probability matrices were estimated for alternative trees using ML and the same program used in Adachi and Hasegawa (1996a).

Table 1. Comparison of log-likelihood of trees for the 45 proteins with the JTT-F model of amino acid substitution

	Without Γ		With Γ		Stripped sites	
Tree	Concatenated	Separate	Concatenated	Separate	Without Γ	With Γ
45 proteins (9,957 sites)					(5,397 sites)	
Tree 1	<−107,958.1>	<−103,229.0>	<−103,923.7>	<−100,665.0>	<−25,647.8>	<−25,370.8>
	(0.8330)	(0.8925)	(0.8240)	(0.8610)	(0.8091)	(0.7182)
Tree 2	−109.8 ± 38.7	−192.8 ± 45.7	−115.5 ± 27.3	−162.7 ± 37.8	−127.2 ± 30.1	−81.9 ± 20.2
	(0.0019)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Tree 3	−29.4 ± 28.6	−40.8 ± 36.0	−17.0 ± 18.0	−30.4 ± 30.8	−19.8 ± 18.9	−7.6 ± 11.1
	(0.1650)	(0.0850)	(0.1724)	(0.0662)	(0.1272)	(0.2146)
Tree 4	−71.5 ± 25.1	−53.9 ± 33.6	−34.8 ± 15.9	−28.9 ± 30.1	−24.5 ± 18.4	−12.1 ± 10.3
	(0.0001)	(0.0225)	(0.0036)	(0.0728)	(0.0637)	(0.0672)
42 proteins (excluding *rpo*B, *rpo*C1, and *rpo*C2, (7,659 sites)					(4,687 sites)	
Tree 1	<−74,516.6>	<−71,019.2>	<−71,831.9>	<−69,323.3>	<−21,815.0>	<−21,617.1>
	(0.7781)	(0.7635)	(0.6886)	(0.7367)	(0.8306)	(0.7554)
Tree 2	−236.3 ± 36.8	−250.8 ± 45.7	−176.0 ± 28.0	−186.4 ± 34.6	−145.2 ± 29.9	−98.1 ± 20.9
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Tree 3	−22.9 ± 23.4	−33.0 ± 27.7	−13.9 ± 14.4	−26.7 ± 20.2	−22.9 ± 17.9	−10.3 ± 10.5
	(0.1070)	(0.0778)	(0.1136)	(0.0464)	(0.0841)	(0.1280)
Tree 4	−26.1 ± 23.0	−23.9 ± 26.7	−10.7 ± 14.8	−14.1 ± 18.7	−22.3 ± 18.0	−10.9 ± 10.4
	(0.1149)	(0.1587)	(0.1978)	(0.2169)	(0.0851)	(0.1166)

The log-likelihoods, ln *L*, of the ML tree are given in angle brackets, and the differences, Δ ln *L*, of alternative trees from that of the ML tree are shown with their SEs (following ±), which were estimated by the formula of Kishino and Hasegawa (1989). Bootstrap proportions in parentheses were estimated by the RELL method (Kishino et al. 1990) with 10^4 replications. "Separate" means that ln *L* of each gene is separately estimated and then is summed up. Tree 1: (*Cyanophora,* ((diatom, rhodophyte), (*Euglena,* chlorophyte))), Tree 2: (*Cyanophora,* (rhodophyte, (diatom, (*Euglena,* chlorophyte)))), Tree 3: ((diatom, rhodophyte), (*Cyanophora,* (*Euglena,* chlorophyte))), Tree 4: ((*Cyanophora,* (diatom, rhodophyte)), (*Euglena,* chlorophyte)).

ML estimates of shape parameter α of the Γ-distribution for Tree 1 (no significant differences for other trees; data not shown) and tree length for Tree 1 (a measure of relative evolutionary rate of each protein) are 0.56 and 2.32 (concatenated sequences), 0.45 and 1.51 (*atpA*), 0.22 and 1.26 (*atpB*), 1.46 and 5.23 (*atpE*), 2.85 and 7.40 (*atpF*), 1.39 and 0.63 (*atpH*), 0.50 and 0.72 (*petB*), 1.02 and 1.58 (*petG*), 0.57 and 0.86 (*psaA*), 0.45 and 0.89 (*psaB*), 0.32 and 0.66 (*psaC*), 0.63 and 2.91 (*psaJ*), 0.29 and 0.57 (*psbA*), 0.69 and 1.02 (*psbB*), 0.60 and 0.95 (*psbC*), 0.35 and 0.63 (*psbD*), 0.57 and 1.05 (*psbE*), 0.62 and 1.07 (*psbF*), 0.56 and 2.25 (*psbH*), 0.70 and 1.67 (*psbI*), 0.99 and 1.93 (*psbJ*), 0.74 and 2.63 (*psbK*), 0.38 and 1.38 (*psbL*), 1.71 and 2.80 (*psbN*), 1.07 and 2.06 (*psbT*), 0.81 and 1.87 (*rpl14*), 0.66 and 2.18 (*rpl*16), 0.76 and 2.71 (*rpl*2), 1.06 and 4.81 (*rpl20*), 1.07 and 4.60 (*rpl22*), 0.55 and 2.17 (*rpl36*), 0.97 and 3.98 (*rpoB*), 0.91 and 3.85 (*rpoC*1), 0.94 and 5.48 (*rpoC*2), 1.33 and 2.94 (*rps*11), 0.38 and 1.18 (*rps*12), 0.67 and 3.57 (*rps*14), 1.11 and 2.95 (*rps*18), 1.00 and 2.66 (*rps*19), 1.23 and 3.56 (*rps*2), 1.01 and 4.18 (*rps*3), 1.03 and 3.28 (*rps*4), 1.45 and 3.18 (*rps*7), 1.04 and 3.90 (*rps*8), 2.10 and 4.24 (*ycf*4), 1.96 and 5.30 (*ycf*9).

Phylogenetic Tree of Proteins in Chloroplast Genomes

Our first task is to estimate the most reliable tree for these chloroplast sequences using both ML and AA LogDet methods. Because estimation of a transition probability matrix will be sensitive to the assumed phylogenetic tree, we will hopefully obtain a reliable tree prior to doing this. It also makes sense to consider how much estimates of the inferred transition matrix vary depending on the tree selected, which we do later for all reasonable trees.

ML Tree Analyses

The *Euglena*/chlorophyte grouping and within-chlorophytes relationships are not biologically contentious and were well supported (e.g., Martin et al. 1998). We therefore fixed the relationships of (*Euglena,* (*Marchantia,* (*Pinus,* (*Nicotiana,* (*Zea, Oryza*))))) within the *Euglena*/chlorophyte clade in our ML analyses. That left 15 possible trees for the four ingroups, *Euglena*/chlorophyte, rhodophyte, diatom, *Cyanophora,* and the outgroup *Synechocystis* to examine.

The ProtML program from MOLPHY (Adachi and Hasegawa 1996b) and the AAML program in PAML (Yang 1997) were applied by assuming site homogeneity and site heterogeneity, respectively, with the JTT-F model, and the results are given in Table 1. As in Martin et al. (1998), the only trees (Fig. 1) we recovered in any of 10,000 RELL bootstrap replicates in this paper were: Tree 1: (*Cyanophora,* ((diatom, rhodophyte), (*Euglena,* chlorophyte))), Tree 2: (*Cyanophora,* (rhodophyte, (diatom, (*Euglena,* chlorophyte)))), Tree 3: ((*diatom,* rhodophyte), (*Cyanophora,* (*Euglena,* chlorophyte))), Tree 4: ((*Cyanophora,* (diatom, rhodophyte)), (*Euglena,* chlorophyte)), and therefore only these four trees are shown in the tables. Although site heterogeneity is taken into account to some extent by using the discrete Γ-distribution, since the lineage specific rates of evolution can differ among different proteins, we also carried out ML esti-

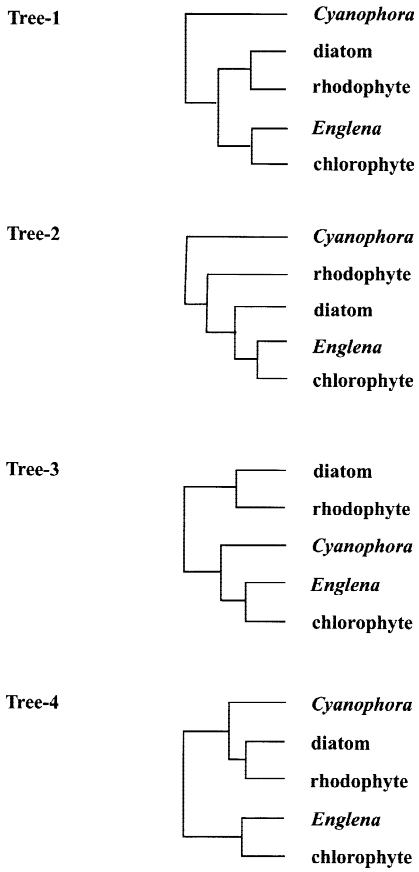

Fig. 1. Candidate plastid phylogenies discussed in this paper (with *Synechocystis* as an outgroup).

mation for each individual protein separately, then summed up the estimated log-likelihoods (ln L) by using the TotalML program in MOLPHY rather than analyze the concatenated sequences. These analyses are also given in Table 1. Note that if the other parts of the model are correct, needing to adopt different relative edge lengths for different genes (which is what summed ML does) can be a sign of non-neutral evolution (Waddell 1995 [p. 469], 2000).

The adequacy of models were compared by using the Akaike Information Criterion, where the AIC score $= -2 \times \ln L + 2 \times$ (number of parameters). The model that minimizes AIC is considered to be the most appropriate model (Akaike 1974). The JTT-F model uses the amino acid frequencies of the data (the number of parameters is 19) and, for a tree with 10 species, 17 branch lengths must be estimated. For the Γ-distribution model, additionally one parameter (shape parameter of the distribution) must be estimated. Therefore, the number of estimated parameters is 37 ($= 17 + 19 + 1$) for each ML analysis with the Γ-distribution versus 36 with the site homogeneity. For Tree 1 and the concatenated analysis of the 45 proteins with the site homogeneity, AIC was 215,988.2, and AIC for the separate analysis was reduced to 209,698.0. Furthermore, the concatenated and separate analyses with the Γ-distribution gave AIC values of 207,921.4 and 204,660.0, respectively. This indicates that the separate analysis approximates the underlying evolutionary process better than the concatenated analysis, which assumes homogeneity of substitution process across genes, and this holds even if the site heterogeneity is taken into account with the Γ-distribution. The separate analyses with the Γ-distribution for each of the individual proteins turned out to best approximate the data given the present models, and the result of the analyses should hopefully be more reliable than those which had worse AIC values. Note, this result is also consistent with non-neutral evolution, although parts of the model need better specification to be more confident of this result (e.g., the form of site rate distribution; Waddell et al. 1997). The result in terms of tree preference does not differ very much for different models. All the analyses prefer Tree 1, and Trees 3 and 4 receive minor support, while Tree 2 is clearly rejected.

As noted in Martin et al. (1998), the subunits of the chloroplast DNA (cpDNA) encoded RNA polymerase *rpoB*, *rpoC*1, and *rpoC*2 favor Tree 2, and the log-likelihood of Tree 1 is lower than those of Tree 2 by 12.3 ± 6.4 (± 1 SE), 2.9 ± 4.5, and 8.5 ± 4.8, respectively, for the JTT-F model with the discrete Γ-distribution. Table 1 also shows the results of the ML analyses for the 42 proteins excluding these subunits. These analyses again support Tree 1, although the BPs for Tree 1 are a little bit lower than they were with 45 proteins, and the BP of Tree 4 is increased.

Naylor and Brown (1997) have recently diagnosed amino acids which frequently substitute as the major cause of erroneous tree estimation when using mtDNA. Therefore, we carried out AAML analysis after converting all valines, leucines, and methionines into isoleucines and all lysines into arginines, but we did not obtain significant differences from the analysis of the original data (log-likelihood differences $\Delta \ln L$ of Trees 2, 3, and 4 from Tree 1 were −86.3 ± 21.3, −10.4 ± 16.5, and −26.8 ± 14.1 for the 45 proteins).

Our site stripping analysis was like that in Waddell et al. (1999) and involved removing all amino acid positions showing any change within the monophyletic group of *Euglena*/land plants. This leaves 5,397 and 4,687 sites for the 45 and 42 proteins, respectively, these generally being an unbiased set of the slowest evolving sites, which are expected to be most robust to misspecification of the ML model (assuming relative rates of sites are similar across the tree). The site stripping ML results again favor Tree 1, exclude Tree 2, and leave Trees 3 and 4 with low levels of support. Interestingly, while the support of Tree 1 decreases for the 42 protein-data relative to the 45 protein-data in the conventional analyses, the support of Tree 1 increases for the 42-protein-data after site stripping analysis. No such increase in the support of Tree 4 was observed.

Table 2. G^2 statistics of amino acid composition stationarity (for the 9,957 sites)

Species	2	3	4	5	6	7	8	9	10
1 *Synechocystis*	302.6	165.5	191.8	702.1	322.6	160.5	166.1	178.3	174.8
2 *Odontella*	—	75.5	53.2	182.1	43.4	115.6	141.4	105.0	108.1
3 *Porphyra*	—	—	35.9	326.6	89.1	73.2	92.5	75.3	75.4
4 *Cyanophora*	—	—	—	326.1	63.0	93.0	97.8	71.9	76.9
5 *Euglena*	—	—	—	—	223.0	406.3	430.6	340.5	342.9
6 *Marchantia*	—	—	—	—	—	220.9	218.1	133.6	150.6
7 *Pinus*	—	—	—	—	—	—	33.9	39.7	37.0
8 *Nicotiana*	—	—	—	—	—	—	—	39.7	39.4
9 *Zea*	—	—	—	—	—	—	—	—	21.2
10 *Oryza*	—	—	—	—	—	—	—	—	—

$\Sigma G^2 = 7{,}363.0$, $\Sigma df = 855$. About 40 (19 d.f.) is quite significant, so all pairs of sequences are clearly different in amino acid composition, except within the last three, which are all higher plants.

Biased AA Composition and LogDet

Using the pairwise test of Waddell et al. (1999) (see also Penny et al. 1999), the amino acid composition is seen to vary significantly between lineages (Table 2), thus violating the stationarity assumption in the ML models used above. This in turn may cause inconsistency of tree estimation (Lockhart et al. 1994; Swofford 1996). In particular, *Euglena* deviates most strongly from other genomes studied here at the level of amino acid composition.

AA LogDet distances (Penny et al. 1999; Waddell et al. 1999) take into account variation of amino acid composition across lineages, when sites evolve uniformly. Their use, combined with NJ, groups the diatom with the *Euglena*/chlorophyte group (Tree 2) with a BP of 77% for the 45 protein concatenation. The proportion of invariant sites estimated by a monophyletic group capture-recapture method (Waddell et al. 1999) was 3,926 out of 9,957 (39.4%). The monophyletic group identified for this method was the *Euglena*/chlorophyte group, which is diverse but not contentious. These inferred invariant sites were then excluded in proportion to the amino acid frequency of the constant sites (Waddell 1995). Analysis of this data set with AA LogDet distances (or the p_{inv}-AA LogDet, e.g., Waddell 1995) favors both the rhodophyte/diatom grouping and *Cyanophora* as sister to the chloroplast sequences (Tree 1) at 55% BP. Thus, with the more appropriate distances, the NJ analyses are in agreement with the ML analyses.

Next, when the three *rpo* proteins are excluded from the analysis, the rhodophyte/diatom grouping is supported with 100% BP by p_{inv}-AA LogDet, and Tree 2 was not recovered in 1,000 bootstrap replications. Although Tree 1 is the NJ tree in this second analysis, the BP for the sister-group relationship of *Cyanophora* to all other plastids (with the cyanobacteria as an outgroup) is only 30.7%, so Tree 1 is not discriminated from Trees 3 and 4.

We also carried out ML analyses of the data sets when the sequence with the most different AA composition, *Euglena,* was excluded (Table 3). While there is no guarantee that the unequal AA composition of the other sequences is not still causing a problem, this approach is an interesting check. In this case, the support for Tree 1 decreased slightly compared to the values shown in Table 1 (except after site stripping of the 45 proteins), and Tree 2 was rejected even more strongly.

Tree 2 Is Quite Unlikely

To make this point clearly, we need to correct an error in a previous paper. In study of Martin et al. (1998), it was reported that Tree 1 was supported by ML analyses of a 11,039-site data (including gapped sites) from the same 45 proteins as studied here, while Trees 2, 3, and 4 were not excluded. It was also claimed that NJ with both Dayhoff (NJ-D) and Kimura (NJ-K) distance estimates, as well as parsimony (MP), gave Tree 1 100% bootstrap support. However, rechecking the results and data revealed that the NJ-D, NJ-K, and MP results reported in Martin et al. (1998) actually correspond to a larger data set of 46 proteins consisting of 11,521 sites, including the *rbcL* sequences from the same genomes. Correct BPs with the 11,039-site and 11,521-site data sets using NJ-D, NJ-K, and MP are shown in Table 4. NJ with the Dayhoff and Kimura models and MP do detect only Tree 1 for the 11,521-site data set. However, when *rbcL* is removed, these three methods support Tree 2, contrary to ML but consistent with distance Hadamard analysis using AA LogDet distances (Lockhart et al. 1999). It seems that excluding *rbcL* removes a strong bias from the 11,521-site data that discriminates against Tree 2, due to shared paralogy of this gene in *Odontella* and *Porphyra.* Also, support for Tree 1 is restored, using NJ-D, NJ-K, and MP on the 9,958-site data set, if gapped sites are excluded.

It was previously shown that Tree 2 is strongly favored by the three *rpo* genes, *rpoB, rpoC*1, and *rpoC*2

Table 3. Comparison of ln *L* of trees excluding *Euglena* with the JTT-F model of amino acid substitution (concatenated sequences)

	All sites		Stripped sites	
	Without Γ	With Γ	Without Γ	With Γ
45 proteins				
	(9,957 sites)		(7,129 sites)	
Tree 1	<−94,154.6>	<−91,230.9>	<−42,485.7>	<−41,726.0>
	(0.5073)	(0.6048)	(0.8330)	(0.7444)
Tree 2	−153.8 ± 39.0	−132.1 ± 26.1	−149.5 ± 33.4	−99.1 ± 21.6
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Tree 3	−3.0 ± 27.4	−4.2 ± 16.8	−25.5 ± 22.7	−11.2 ± 13.1
	(0.4808)	(0.3884)	(0.0981)	(0.1660)
Tree 4	−38.8 ± 24.4	−25.3 ± 14.1	−25.2 ± 22.6	−14.0 ± 12.6
	(0.0119)	(0.0068)	(0.0689)	(0.0896)
42 proteins excluding *rpo*B, *rpo*C1, and *rpo*C2				
	(7,659 sites)		(5,814 sites)	
Tree 1	<−65,589.7>	<−63,669.9>	<−32,373.9>	<−31,914.1>
	(0.4067)	(0.4474)	(0.5948)	(0.5434)
Tree 2	−269.4 ± 38.4	−188.6 ± 27.9	−211.7 ± 33.4	−138.7 ± 23.6
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Tree 3	−3.6 ± 23.0	−3.7 ± 14.1	−22.2 ± 20.9	−11.0 ± 12.2
	(0.2413)	(0.2774)	(0.0701)	(0.0972)
Tree 4	−1.6 ± 22.9	−3.6 ± 14.0	−8.6 ± 22.1	−3.4 ± 13.3
	(0.3520)	(0.2752)	(0.3351)	(0.3594)

(Martin et al. 1998). Among the proteins common to these chloroplast genomes, the three *rpo*'s belong to the most highly variable family (Goremykin et al. 1997). Furthermore, due to the presence of numerous internally gapped regions, the *rpo*-proteins contain many sections of uncertain alignment. For example, in the 11,039-site alignment there are a total of 1,082 gapped sites, of which 877 (81%) are found in the three *rpo* genes. In the present paper, gapped sites are excluded from the analyses, and the amount of data remaining, 9957 sites, is still large. All of the ML analyses of this paper exclude Tree 2, regardless of whether the *rpo* genes are included (45 protein data) or excluded (42 protein data); a Γ-distribution of rate heterogeneity across sites was considered or not; only the highly conservative site strippings were considered; the JTT-F or cpREV (see below) substitution matrix is used.

In summary, our phylogentic analyses, both ML and AA LogDet, cannot exclude Trees 3 and 4, but strongly discriminate against Tree 2. The highly variable *rpo* genes support Tree 2. Apparently, this signal leads to strong bootstrap support for Tree 2 with simple substitution models, overriding the signal contained in the remaining 42 genes. Bootstrap estimates of support can be susceptible to bias in a model-dependent manner; the results here suggest that the LogDet and ML models are less vulnerable to such bias than the Dayhoff and Kimura distances and MP in the case of this data. Notably, strong evidence favoring the position of *Odontella* in Tree 1 over Tree 2 is found in the chloroplast operon organization (McFadden et al. 1997; Leitsch et al. 1999). If this data is correct, then the results here underscore the importance of realistic substitution models that can cope with composition bias (Lockhart et al. 1994, 1999; Waddell 1995; Hasegawa and Adachi 1996; Waddell et al. 1999) in use in phylogenetic inference. In all cases, Trees 2 and 3 were also still viable, due to some difficulty in locating where the outgroup should join the ingroup tree.

The cpREV Model of Amino Acid Substitution

Since we cannot be certain of the correct tree, we estimated transition probability matrices for Trees 1, 2, 3, and 4, respectively, using the 9,957-sites data. Table 5 (also **www.evol.ism.ac.jp**) gives the transition probability matrix estimated assuming Tree 1, which we refer to as cpREV.

Table 6 shows the result of the ML analysis of the 45 proteins (9,957 sites) with the cpREV matrix. Although Trees 3 and 4 cannot be excluded, the differences of the result due to using cpREV are minor. While the concatenated analysis of the 45 proteins using cpREV and the Γ-distribution gives a maximum ln *L* of −103,141.2 (Table 6), those by the JTT-F, Dayhoff-F, and mtREV-F give −103,923.7 (lower by 782.5 ± 45.6 than that by the cpREV), −104,765.3 (lower by 1,624.1 ± 66.3) and −105,612.4 (lower by 2,471.1 ± 97.9), respectively. These values show that the JTT model of nuclear-encoded proteins approximates the amino acid substitutions of chloroplast-encoded proteins much better than the mtREV model does. This is probably due to the codes used in the nuclear and chloroplast systems being the universal code and quite distinct from that of studied mitochondria.

We also ran the analyses using the matrices estimated for Trees 2, 3, and 4, with results shown in Table 7. The

Table 4. Bootstrap support for Tree 1 versus Tree 2

		BP (%)	
		Tree 1	Tree 2
46 proteins			
11,521 sites			
	ML	78.0	0.0
	NJ-D	100	0
	NJ-K	100	0
	MP	100	0
45 proteins excluding *rbc*L			
11,039 sites			
	ML	82.4	6.1
	NJ-D	3	97
	NJ-K	12	88
	MP	0	100
Excluding gaps			
9,958 sites			
	ML	83.7	0.2
	NJ-D	8	92
	NJ-K	18	82
	MP	36.5	63.5
42 proteins excluding *rpo*B, *rpo*C1, *rpo*C2			
7,864 sites			
	ML	68.8	0.0
	NJ-D	100	0
	NJ-K	100	0
	MP	90.5	5.5
Excluding gaps			
7,660 sites			
	ML	74.5	0.0
	NJ-D	99	0
	NJ-K	98	0
	MP	89.5	0

Bootstrap proportions for Tree 1 and Tree 2 with neighbor-joining using the Kimura (NJ-K) or Dayhoff (NJ-D) or maximum parsimony (MP). 100 samples were used.

results are similar to the previous ones. In particular, even if the optimal transition matrix for Tree 2 is used, Tree 2 is again rejected with a high statistical significance. This holds also if site heterogeneity is taken into account. There is only a slight favoritism for selecting the tree that rate matrix was estimated on. This result gives us confidence that cpREV will be a useful model, even if Tree 1 turns out not to be correct. These fairly mild effects are probably due to the difference between these trees being a few short internal edges; with more taxa, and longer edges, a bias due to which tree is assumed when estimating the rate matrix should be watched out for. For those who are particularly concerned at such effects, we offer a rate matrix that is assessed over all feasible trees; i.e., a weighted sum of rate matrices, where the RELL proportion of each feasible tree is used as the weight (we call this cpREV*, available at **www.evol.ism.ac.jp**). In particular, the weights used were (0.75, 0.13, 0.12) or the RELL bootstrap proportions from the site-stripped data for the 42 proteins assuming the Γ-JJT-F model (perhaps the most realisitic assessment of support for the different trees) (Table 1).

Note, the average amino acid frequencies for the 45 chloroplast proteins studied here are more similar to nuclear proteins (compiled by Jones et al. 1992) than to mitochondrial proteins of vertebrates (mtREV; Adachi and Hasegawa 1996a), with the exceptions of cysteine and isoleucine (data not shown). A lower frequency of cysteine in mitochondrial and chloroplastic proteins relative to that of nuclear proteins is probably due to many organelle-encoded proteins being trans-membrane, so the disulfide bonds common in globular proteins are relatively rare.

Since *rbcL* sequences were not used in constructing the cpREV, they can be used to test whether the application of cpREV to the *rbcL* data significantly improves the likelihood over transition matrices estimated using nuclear and mtDNA encoded genes. To do this, we examined a data set of 22 *rbcL* sequences from 6 ferns, 2 byophytes, 2 algae, and 11 seed plants, including 3 angiosperms, 3 gnetales, 2 conifers, 2 cycads, and *Ginkgo* (data available via FTP from **134.169.70.80/ftp/pub/incoming/adachi/rbcl.data**). We also wished to look at the effect of cpREV with respect to the branching order of seed plant groups, since this is currently a highly debated topic. In particular, the relationship of the gnetophytes to the angiosperms is controversial because the anthophyte hypothesis holds Gnetales to be the sister group of angiosperms (reviewed in Crane et al. 1995), although recent molecular data are equivocal on this view (e.g., Crepet 1998, Doyle 1998), and the latest tend to reject it (Hansen et al. 1999; Samigullin et al. 1999; Winter et al. 1999). Specifically, we examined all 105 rooted trees for the five seed plant clades assuming identical site rates. Then, the best tree among those 105 was compared using the Γ-distribution to the best tree that contained the angiosperm–Gnetales clade. The best tree contained the branching order (((Gnetales, (conifers, *Ginkgo*), cycads), angiosperms), outgroups) with ln L of −3,364.94 (Dayhoff-F), −3,340.02 (JTT-F), −3,436.10 (mtREV-F), and −3,327.23 (cpREV-F). The best tree grouping angiosperms and Gnetales as sisters showed (((*Ginkgo,* conifers), (cycads, (Gnetales, angiosperms))), outgroups) (−2.09 ± 7.91 for Dayhoff-F, −2.27 ± 7.45 for JTT-F, −0.05 ± 7.11 for cpREV-F), although the standard error of Δ ln L was much larger than Δ ln L. Thus, although the *rbcL* amino acid sequences were insufficient to discriminate between these topologies, cpREV-F does better fit the data than JTT-F. Interestingly, the rate matrix cpDNA*-F (ln L = −3,326.98) did slightly better (0.25 ln$L$ units) than cpREV-F. This suggests that estimating a weighted rate matrix using all the feasible trees may offer advantages in predictive power over estimating it on the ML tree only.

Last, one might wonder why Trees 3 and 4 cannot be excluded even using whole chloroplast genomes. To in-

Table 5. Transition probability matrix $P_{ij}(\times 10^6)$ of the amino acid *i* being replaced by the amino acid *j* during a time interval of one substitution per 100 amino acids (1PAM) for the cpREV model (estimated for Tree 1), and average amino acid frequencies π of the cpDNA-encoded proteins

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
Ala	990,888	125	180	124	115	115	469	1,061	31	228
Arg	152	991,038	281	32	141	1,265	150	389	335	209
Asn	332	426	985,292	3,108	93	558	995	1,040	656	258
Asp	253	53	3,435	989,791	2	294	3,459	689	156	17
Cys	965	968	419	8	989,592	8	11	484	207	432
Gln	228	2,059	600	286	2	987,841	2,926	214	593	142
Glu	717	188	824	2,590	2	2,255	988,094	605	77	228
Gly	958	288	508	305	52	97	357	995,946	9	62
His	96	845	1,091	235	76	919	154	31	994,224	44
Ile	214	161	131	7	48	67	140	64	13	986,820
Leu	284	240	89	7	68	207	78	32	31	2,667
Lys	342	5,255	1,878	294	9	2,389	2,459	421	144	528
Met	267	148	48	33	27	146	107	33	4	2,707
Phe	99	63	76	16	125	8	136	40	60	698
Pro	706	104	136	120	49	235	175	46	71	180
Ser	3,491	456	1,614	416	400	287	535	1,100	142	332
Thr	1,924	372	1,081	189	99	176	349	149	15	1,589
Trp	21	271	31	12	75	38	59	130	32	65
Tyr	82	381	586	199	252	284	135	17	922	138
Val	1,387	110	65	52	101	40	188	145	11	7,299
π	0.0756	0.0621	0.0410	0.0371	0.0091	0.0382	0.0495	0.0838	0.0246	0.0806

Table 6. Comparison of ln *L* of trees for the 45 proteins with the cpREV model of amino acid substitution

	Without Γ		With Γ	
	Concatenated	Separate	Concatenated	Separate
Tree 1	<−106,507.1> (0.9355)	<−100,731.2> (0.8432)	<−103,141.2> (0.8996)	<−100,244.8> (0.9243)
Tree 2	−106.0 ± 37.1 (0.0025)	−165.0 ± 38.0 (0.0000)	−113.1 ± 27.4 (0.0000)	−163.1 ± 36.3 (0.0000)
Tree 3	−42.1 ± 27.9 (0.0619)	−28.4 ± 30.8 (0.0764)	−24.1 ± 18.2 (0.0972)	−38.3 ± 31.0 (0.0328)
Tree 4	−77.4 ± 24.9 (0.0001)	−38.3 ± 27.3 (0.0804)	−39.2 ± 16.4 (0.0032)	−35.8 ± 29.9 (0.0429)

Table 7. Analyses of the concatenated sequences with transition matrices estimated for the respective trees (with the site homogeneity)

	Matrix for				
	Tree 1	Tree 2	Tree 3	Tree 4	Respective trees
Tree 1	<−106,507.1> (0.9355)	<−106,511.2> (0.9354)	<−106,509.4> (0.9142)	<−106,509.4> (0.9242)	<−106,507.1> (0.9261)
Tree 2	−106.0 ± 37.1 (0.0025)	−97.9 ± 36.9 (0.0044)	−106.6 ± 37.1 (0.0023)	−107.1 ± 37.1 (0.0022)	−102.0 ± 36.9 (0.0036)
Tree 3	−42.1 ± 27.9 (0.0619)	−42.6 ± 27.9 (0.0601)	−37.6 ± 27.8 (0.0833)	−39.1 ± 27.6 (0.0732)	−39.9 ± 27.8 (0.0701)
Tree 4	−77.4 ± 24.9 (0.0001)	−78.3 ± 24.9 (0.0001)	−74.4 ± 24.6 (0.0002)	−72.9 ± 24.7 (0.0004)	−75.2 ± 24.8 (0.0002)

vestigate further whether the failure in resolving among Trees 1, 3, and 4 is due to some complexity in substitution process or whether it might be due simply to the short interval separating relevant branchings, we carried out a parametric bootstrap simulation; i.e., 10 data sets of 9,957-site lengths were generated with the site rate homogeneity and the cpREV matrix, assuming Tree 1 and the branch lengths estimated by ProtML from the real data (i.e., concatenated sequences without Γ in Table 6). These were then analyzed with the same model (Table 8). Only two of the simulated data sets could discriminate Trees 3 and 4 from Tree 1, suggesting that branchings among the diatom/rhodophyte, the *Euglena*/chlorophyte, and the *Cyanophora* lineages might have occurred within a very short time period, perhaps too short to permit the accumulation of sufficient phyloge-

Table 5. Extended

Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
380	227	78	66	402	2,871	1,381	5	33	1,211
391	4,265	52	51	72	457	326	79	188	117
219	2,310	26	94	143	2,450	1,433	13	439	105
20	400	19	22	140	699	277	6	164	94
762	50	67	699	235	2,742	597	150	854	741
550	3,155	84	10	265	469	251	18	228	69
159	2,503	47	139	153	673	383	21	83	251
39	253	8	24	24	816	96	28	6	115
128	296	4	123	125	360	35	23	1,150	31
3,346	330	740	438	96	257	1,071	14	52	5,981
992,087	209	566	1,219	180	609	163	55	111	1,089
420	983,023	81	70	248	1,024	945	3	145	312
2,597	185	991,752	316	82	112	665	29	126	604
2,436	69	137	993,319	36	575	154	161	1,381	400
423	289	42	42	995,456	1418	270	17	57	154
989	829	39	467	983	985,156	2,209	25	304	213
304	876	270	143	214	2,531	988,702	10	42	955
306	10	36	451	40	86	30	998,082	202	13
367	238	90	2,278	80	617	75	119	992,982	149
1,668	238	201	307	101	201	785	3	69	987,018
0.1011	0.0504	0.0220	0.0506	0.0431	0.0622	0.0543	0.0181	0.0307	0.0660

netic resolution from the standpoint of the available data. If so, why an apparent change in operon structure might have occurred in this time is unclear.

Discussion

The cpREV substitution matrix presented here appears to better approximate the frequencies of amino acid transitions in chloroplast protein data, although the difference to the JTT-F model is not as great as in the case of the mtREV matrix for mitochondrial proteins, where a very significant increase of likelihood is obtained (Adachi and Hasegawa 1996a), probably due to the identity of nuclear and chloroplast genetic codes. Tree 1, used to estimate the cpREV matrix, seems quite robust from the standpoint of the variations of analysis investigated here.

When nonstationarity of amino acid or nucleotide composition is expected, a number of different tests are possible. The traditional "Block" test (e.g., in PAUP*, Swofford 1999) is incorrect, partly because it includes both invariant and by chance unvaried sites (Waddell 1995, Waddell et al. 1999). In its place may be used the pairwise tests of Waddell et al. (1999) and Penny et al. (1999). Their use here clearly showed that all pairs of species were significantly different except those of the higher plants, which here can only mean considerable nonstationarity. This result is very similar to that obtained for mammalian mtDNA (Waddell et al. 1999; Penny et al. 1999), suggesting that the use of the Block test has generally been hiding significant nonstationarity from biologists. The generalized least squares (GLS) test is a more formally correct test for a set of taxa (Rzhetsky and Nei 1995), but it would be computationally very expensive for this data, is unavailable as far as we are aware, and would not immediately identify which taxa were different to which others. For the loss of statistical power relative to the GLS test, the pairwise test allows the user to get a good feeling for the structure of nonstationarity in their data.

All ML analyses reported here favor Tree 1. This is true (1) for ProtML in concatenate and separate analyses, (2) using the JTT-F or the cpREV model, (3) taking between-site variation of the substitution rate into account assuming a gamma site rate distribution, (4) taking into account the substitution patterns of highly variable amino acids (Naylor and Brown 1997), (5) excluding or including *Euglena* (which has a highly aberrant amino acid composition), (6) excluding highly variable positions by site-stripping (Waddell et al. 1999), and (7) excluding the highly variable *rpo* proteins. The NJ analyses using the LogDet model for estimating distances, which takes biases in the data into account, also support Tree 1. Tree 1 is furthermore supported by independent data from the comparison of chloroplast operon organization (Leitsch et al. 1999).

Despite this, the data still do not permit us to fully exclude Tree 3 and Tree 4. Given the amount of data analyzed here and the difficulties that we have encountered in statistically discriminating Tree 1 from Tree 3 and Tree 4, it seems that even more data is necessary to completely resolve this topology and, by inference, similarly deep branching patterns in early evolution. For plastid genomes, there is not very much more molecular data available (Martin et al. 1998). If *Euglena* were to be entirely excluded, nine more proteins are common to the remaining taxa could be included in the data: *ccsA* (353), *petA* (320), *petD* (160), *petL* (31), *rpl*33 (65), *psaI* (36),

Table 8. ProtML analyses of simulation data of 9,957 sites generated with the cpREV model for Tree 1 with the branch lengths estimated for the real data by the ProtML. Bootstrap proportions in parentheses were estimated by the RELL method (Kishino et al. 1990) with 10^4 replications

	Tree 1	Tree 2	Tree 3	Tree 4
Real data	<−106,507.1>	−106.0 ± 37.1	−42.1 ± 27.9	−77.4 ± 24.9
	(0.9310)	(0.0017)	(0.0672)	(0.0001)
Simulation data				
simu-1	<−30,428.8>	−55.1 ± 21.5	−6.8 ± 6.0	−6.8 ± 6.0
	(0.8423)	(0.0012)	(0.0597)	(0.0968)
simu-2	<−30,373.8>	−14.7 ± 10.0	−16.4 ± 11.0	−16.4 ± 11.0
	(0.9023)	(0.0477)	(0.0152)	(0.0348)
simu-3	<−30,300.5>	−29.4 ± 15.4	−4.9 ± 5.8	−4.9 ± 5.8
	(0.5876)	(0.0091)	(0.1755)	(0.2278)
simu-4	<−30,467.3>	−19.7 ± 11.6	−2.8 ± 3.6	−2.8 ± 3.6
	(0.5961)	(0.0170)	(0.2537)	(0.1332)
simu-5	<−30,438.5>	−41.5 ± 17.9	−13.4 ± 10.7	−13.4 ± 10.7
	(0.8787)	(0.0049)	(0.0523)	(0.0641)
simu-6	<−30,453.5>	−20.2 ± 13.4	−12.1 ± 9.8	−12.1 ± 9.8
	(0.8410)	(0.0456)	(0.0466)	(0.0668)
simu-7	<−30,580.1>	−32.1 ± 16.6	−41.6 ± 17.8	−41.6 ± 17.8
	(0.9877)	(0.0098)	(0.0012)	(0.0013)
simu-8	<−30,517.3>	−35.7 ± 16.6	−16.5 ± 11.1	−16.5 ± 11.1
	(0.9539)	(0.0012)	(0.0146)	(0.0303)
simu-9	<−30,243.3>	−47.3 ± 20.0	−33.5 ± 15.7	−33.5 ± 15.7
	(0.9904)	(0.0022)	(0.0033)	(0.0041)
simu-10	<−30,363.5>	−32.6 ± 15.2	−8.9 ± 7.5	−8.9 ± 7.5
	(0.8313)	(0.0045)	(0.0836)	(0.0806)

rpoA (311), *ycf*3 (173), and *ycf*6 (29), providing about 1,478 total additional amino acid sites (roughly 15% more). Another tact might be tracing chloroplast genes that have moved to the nucleus with care to discriminate whether their phylogeny is homologous with that of the common chloroplast genes or not. This could allow nearly all the genes found in all these chloroplasts to be used.

The cpREV substitution matrix should be a useful tool for analysis of data from protein-coding regions of chloroplast DNA because it serves the purpose of providing more realistic models of amino acid substitution for chloroplast-encoded proteins. Even with very large amounts of data and using a number of reasonably sophisticated models, the problem of resolving the relatively deep branches of plastid phylogeny is surprisingly severe. This suggests that resolving even deeper branches of phylogeny with molecular data presents a formidable task that will require realistic models of sequence evolution, information from many genes, and (ideally) supplemental independent data like gene order.

Acknowledgments. We thank two anonymous referees for helpful comments. This work was supported by grants from the Ministry of Education, Science, Sports and Culture of Japan (M.H. and P.J.W.), JSPS (P.J.W.), Marsden Fund of New Zealand (P.J.W.), and the Deutsche Forschungsgemeinschaft (W.M.).

References

Adachi J, Hasegawa M (1996a) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Adachi J, Hasegawa M (1996b) MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr Inst Stat Math 28:1–50

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr AC-19:716–723

Barry D, Hartigan J (1987) Asynchronous distance between homologous DNA sequences. Biometrics 43:261–276

Bhattacharya D (1997) An introduction to algal phylogeny and phylogenetic methods. In: Bhattacharya D (ed) Origins of algae and their plastids. Springer Verlag, New York, p 1

Cao Y, Janke A, Waddell P, Westerman M, Takenaka O, Murata S, Okada N, Pääbo S, Hasegawa M (1998a) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol 47:307–322

Cao Y, Waddeil P, Okada N, Hasegawa M (1998b) The complete mitochondrial DNA sequence of the shark *Mustelus manazo:* evaluating rooting contradictions to living bony vertebrates. Mol Biol Evol 15:1637–1646

Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature 374:27–33

Crepet W (1998) The abominable mystery. Science 282:1653–1654

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol. 5, suppl 3. National Biomedical Research Foundation, Washington DC, pp 345–352

Delwiche C, Palmer J (1997) The origin of plastids and their spread via secondary symbiosis. In: Bhattacharya D (ed) Origins of algae and their plastids. Springer Verlag, New York, pp 53–86

Doyle JA (1998) Molecules, morphology, fossils and the relationship of angiosperms and Gnetales. Mol Phyl Evol 9:448–462

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Goremykin V, Hansmann S, Martin W (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast ge-

nomes: revised molecular estimates of two seed plant divergence times. Plant Syst Evol 206:337–351

Hansen A, Hansmann S, Samigullin T, Antonov A, Martin W (1999) Gnetum and the angiosperms: molecular evidence that their shared morphological characters are convergent, rather than homologous. Mol Biol Evol 16:1006–1009

Hasegawa M, Adachi J (1996) Phylogenetic position of cetaceans relative to artiodactyls: reanalysis of mitochondrial and nuclear sequences. Mol Biol Evol 13:710–717

Hasegawa M, Kishino H (1994) Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. Mol Biol Evol 11:142–145

Janke A, Xu X, Arnason U (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia and Eutheria. Proc Natl Acad Sci USA 94:1276–1281

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

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

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

Lake J (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91:1455–1459

Leitsch C, Kowallik K, Douglas S (1999) The *atpA* gene cluster of a cryptomonad, *Guillardia theta:* a piece in the puzzle of chloroplast genome evolution. J Phycol (in press)

Lockhart P, Howe C, Barbrook A, Larkum A, Penny D (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol 16:573–576

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

Maier R, Neckermann K, Igloi G, Kössel H (1995) Complete sequence of maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol, 251:614–628

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

McFadden G, Waller R, Reith M, Lang-Unnasch N (1997) Plastids in apicocomplexan parasites. In: Bhattacharya D (ed) Origins of algae and their plastids. Springer Verlag, New York, pp 261–287

Muse S, Gaut B (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Naylor G, Brown W (1997) Structural biology and phylogenetic estimation. Nature 388:527–528

Penny D, Hasegawa M, Waddell P, Hendy M (1999) Mammalian evolution: Timing and implications from using the LogDeterminant transform for proteins of differing amino acid composition. Syst Biol 48:76–93

Perna N, Kocher T (1995) Unequal base frequencies and the estimation of substitution rates. Mol Biol Evol 12:359–361

Reeves JH (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol 35:17–31

Rzhetsky A, Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol 12:131–151

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Samigullin TH, Martin W, Troitsky AV, Antonov AS (1999) Molecular data from the chloroplast rpoC1 gene suggest a deep and distinct dichotomy of contemporary spermatophytes into two monophylums: gymnosperms (including Gnetales) and angiosperms. J Mol Evol 49:310–315

Soltis P, Soltis D, Doyle J (1992) Molecular systematics of plants. Chapman & Hall, New York

Swofford D (1996) PAUP*. National Museum of Natural History, Smithsonian Institution, Washington, DC

Swofford DL (1999) PAUP*—phylogenetic analysis using parsimony (*and other methods), version 4.0b1. Sinauer Associates, Sunderland, MA

Waddell PJ (1995) Statistical methods of phylogenetic analysis: including Hadamard conjugations, LogDet transforms, and maximum likelihood. PhD diss., Massey University

Waddell PJ (2000) Phylogenetic analysis of sequence data. In: Baxevanis AD, Francis Oullette BF (eds) Bioinformatics, 2nd ed. Wiley, New York (in press)

Waddell PJ, Steel MA (1997) General time reversible distances with unequal rates across sites: mixing and inverse Gaussian distributions with invariant sites. Mol Phyl Evol 8:398–414

Waddell PJ, Penny D, Moore, T (1997) Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol Phyl Evol 8:33–50

Waddell P, Cao Y, Hauf J, Hasegawa M (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid–invariant sites–LogDet 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

Winter K-U, Becker A, Muenster T, Kim JT, Saedler H, Theissen G (1999) MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc Natl Acad Sci USA 96:7342–7347

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. CABIOS 13:555–556

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600–1611

Zardoya R, Cao Y, Hasegawa M, Meyer A (1998) Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data. Mol Biol Evol 15:506–517