J Mol Evol (1990) 31:151–160

Journal of Molecular Evolution

Maximum Likelihood Inference of Protein Phylogeny and the Origin of Chloroplasts

Hirohisa Kishino,[1] Takashi Miyata,[2] and Masami Hasegawa[1]

[1] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan
[2] Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan

Summary. A maximum likelihood method for inferring protein phylogeny was developed. It is based on a Markov model that takes into account the unequal transition probabilities among pairs of amino acids and does not assume constancy of rate among different lineages. Therefore, this method is expected to be powerful in inferring phylogeny among distantly related proteins, either orthologous or paralogous, where the evolutionary rate may deviate from constancy. Not only amino acid substitutions but also insertion/deletion events during evolution were incorporated into the Markov model. A simple method for estimating a bootstrap probability for the maximum likelihood tree among alternatives without performing a maximum likelihood estimation for each resampled data set was developed. These methods were applied to amino acid sequence data of a photosynthetic membrane protein, *psb*A, from photosystem II, and the phylogeny of this protein was discussed in relation to the origin of chloroplasts.

Key words: Evolutionary tree — Amino acid sequence — Insertion/deletion — Bootstrap probability — *psb*A — *Prochlorothrix*

Introduction

As DNA and protein sequence data accumulate, there is an increasing demand for statistical methods to infer evolutionary trees from them. This approach, called molecular phylogenetics, provides us with an objective basis for clarifying phylogenetic relationships among organisms, circumventing the measure of subjectivity involved in the traditional morphological approach. Furthermore, phylogenetic analysis of paralogous genes provides a basis for studying the mechanism of the evolution of genes with new functions.

A maximum likelihood method for inferring trees from DNA sequence data was developed by Felsenstein (1981). Because this method has a sound statistical basis (Felsenstein 1983a,b; Kishino and Hasegawa 1989), we have used it extensively in making inferences about evolutionary relationships among organisms (Hasegawa et al. 1985, 1988; Hasegawa and Kishino 1989; Kishino and Hasegawa 1989). One of the desirable properties of Felsenstein's method is that, as it does not impose any constraint on the constancy of the evolutionary rate, it can infer a correct tree even if the evolutionary rate differs considerably among lineages (Hasegawa and Yano 1984). Because the evolutionary rate can differ sometimes among taxonomic units (Kikuno et al. 1985; Wu and Li 1985; Britten 1986), constancy of the rate should not be assumed in advance in inferring the tree topology, and Felsenstein's method is recommended in this respect.

Another method of maximum likelihood for inferring DNA and protein trees was developed by Bishop and Friday (1985), and they were the first to apply maximum likelihood to real phylogenetic problems involving protein trees (Bishop and Friday 1987). Contrary to Felsenstein, Bishop and Friday assumed constancy of evolutionary rate among lineages. Furthermore, they assumed equal transition probability among different pairs of amino acids, which seems not to be the case in evolution.

Offprint requests to: M. Hasegawa

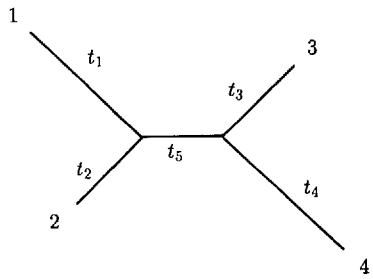

Fig. 1. The unrooted tree used in the discussion of estimating the lengths of branches from amino acid sequences.

Here, we develop a maximum likelihood method for inferring protein trees that takes into account the unequal transition probabilities among pairs of amino acids by using an empirical transition matrix compiled by Dayhoff et al. (1978). It does not assume constancy of the rate among lineages. It takes into account not only amino acid substitutions but also insertion/deletion events in amino acid sequences. Furthermore, we develop a method for estimating bootstrap probabilities of the highest likelihood among alternative trees without performing a maximum likelihood estimation for each resampled data set. The methods are applied to amino acid sequence data of a photosynthetic membrane protein, *psb*A, from photosystem II, and the phylogenetic place of *Prochlorothrix* is discussed in relation to the origin of chloroplasts.

Methods

Markov Model of Amino Acid Substitutions. The Markov process of amino acid sequence evolution is represented by a transition probability matrix of 20 × 20 dimension. If transition probabilities are equal among different pairs of amino acids, the number of amino acid substitutions per site between the pth and qth sequences, $\tilde{D}_{pq}$, is estimated by

$$\tilde{D}_{pq} = -\frac{19}{20}\log\left(1 - \frac{20D_{pq}}{19n}\right) \tag{1}$$

where n is the length of the sequence, and D_{pq} is the number of amino acid differences between p and q. In the case of an unrooted tree for four operational taxonomic units (OTUs) as shown in Fig. 1, numbers of amino acid substitutions along the five branches, t_1, t_2, t_3, t_4, and t_5, can be estimated by a least squares method (Chakraborty 1977) that minimizes

$$\begin{aligned} S = {} & (\tilde{D}_{12} - t_1 - t_2)^2 + (\tilde{D}_{13} - t_1 - t_3 - t_5)^2 \\ & + (\tilde{D}_{14} - t_1 - t_4 - t_5)^2 + (\tilde{D}_{23} - t_2 - t_3 - t_5)^2 \\ & + (\tilde{D}_{24} - t_2 - t_4 - t_5)^2 + (\tilde{D}_{34} - t_3 - t_4)^2 \end{aligned} \tag{2}$$

However, when the transition probability differs among pairs of amino acids, as is the case in the actual process of protein evolution, a model of amino acid substitution contains too many parameters to be estimated, and the complexity of the problem increases tremendously.

In this work, we simplify the problem by introducing an empirical transition matrix compiled by Dayhoff and her coworkers (Dayhoff et al. 1978), who have shown that an amino acid in a protein is replaced more often by a physicochemically similar amino acid than expected under equal transition probability. This observation is consistent with the neutral theory (Kimura 1983). Instead of estimating the transition matrix from the data, we shall use the average transition matrix of Dayhoff, R, as a given one, and we shall estimate $\theta \equiv (t_1, \ldots, t_5)$ by a maximum likelihood. From 71 groups of closely related protein sequences, Dayhoff et al. counted relative transition frequencies A_{ij} ($i, j = 1, 2, \ldots, 20$) between amino acids i and j among the total of 1572 changes, where $A_{ij} = A_{ji}$ and $A_{ii} = 0$ (Fig. 80 in Dayhoff et al. 1978). The fraction of transitions to j among substitutions of amino acid i is

$$B_{ij} = \frac{A_{ij}}{\sum_{k=1}^{20} A_{ik}}$$

They defined relative mutability of amino acid i, m_i, by a ratio of the number of times that the amino acid i has changed to the number of times that it has occurred in the sequences (Table 21 in Dayhoff et al. 1978). A transition probability in a short time interval is given by

$$M_{ij} = \begin{cases} \delta m_i B_{ij} & (i \neq j) \\ 1 - \delta m_i & (i = j) \end{cases} \tag{3}$$

where δ is a constant that determines a unit time interval. When δ is small, the probability that substitution occurs more than twice is negligible, and the number of substitutions in a unit time is given by

$$1 - \sum_{i=1}^{20} \pi_i M_{ii} = \delta \sum_{i=1}^{20} \pi_i m_i$$

where π_i ($i = 1, 2, \ldots, 20$) is the composition of amino acid i (Table 22 in Dayhoff et al. 1978). The process presented by M_{ij} is time reversible.

Neither rate constancy nor constraint on the evolutionary rate is assumed in this analysis. A unit of time is chosen for each branch so that one amino acid substitution occurs per 100 amino acids. When evolutionary rate differs among lineages, a unit of time correspondingly differs among different branches. Because we deal with an unrooted tree and do not assume the constancy of the rate, the length of the ith branch, t_i, is taken as a number of amino acid substitutions per 100 amino acids rather than an absolute time. Let ρ_i and $\mathbf{u}_i$ ($i = 1, 2, \ldots, 20$) be an eigenvalue and an eigenvector of M. Further let

$$\lambda_i = \frac{0.01}{\delta \sum_{i=1}^{20} \pi_i m_i} \log \rho_i \qquad i = 1, \ldots, 20 \tag{4}$$

and

$$U = (\mathbf{u}_1, \ldots, \mathbf{u}_{20}) \tag{5}$$

Then we have

$$R = U \begin{bmatrix} \lambda_1 & & & 0 \\ & \cdot & & \\ & & \cdot & \\ 0 & & & \lambda_{20} \end{bmatrix} U^{-1} \tag{6}$$

Transition probability matrix after an arbitrary time t is given by

$$P(t) = e^{tR} \tag{7}$$

and its component is written as

$$P_{ij}(t) = \sum_{k=1}^{20} c_{ijk} e^{t\lambda_k} \qquad i, j = 1, \ldots, 20 \tag{8}$$

where c_{ijk} is a function of U and U^{-1}.

Maximum Likelihood Procedure. The amino acid sequence data of length n from four species can be represented as follows:

Species 1: $X_{11}\, X_{12}\, X_{13} \ldots X_{1n}$
Species 2: $X_{21}\, X_{22}\, X_{23} \ldots X_{2n}$
Species 3: $X_{31}\, X_{32}\, X_{33} \ldots X_{3n}$
Species 4: $X_{41}\, X_{42}\, X_{43} \ldots X_{4n}$

We write the whole data set as $\mathbf{X}$ and the value of the hth site $(X_{1h}, X_{2h}, X_{3h}, X_{4h})^T$ (a superscript T denotes a transposed vector) as $\mathbf{X}_h$. We assume that each amino acid site evolves independently and identically with others. A probability of occupying amino acids x_1, x_2, x_3, and x_4 at a site in species 1, 2, 3, and 4, respectively, is given by

$$f(x_1, x_2, x_3, x_4 | \theta) = \sum_{i=1}^{20} \left[\pi_i P_{ix_1}(t_1) P_{ix_2}(t_2) \times \sum_{j=1}^{20} P_{ij}(t_5) P_{jx_3}(t_3) P_{jx_4}(t_4) \right] \quad (9)$$

The log-likelihood is

$$l(\theta | \mathbf{X}) = \sum_{h=1}^{n} \log f(\mathbf{X}_h | \theta) \quad (10)$$

We can obtain the maximum likelihood estimate of θ through Newton's method, in which calculations of ∇l and $\nabla\nabla^T l$ are necessary and we have

$$\frac{d}{dt} P_{ij}(t) = \sum_{k=1}^{20} c_{ijk} \lambda_k e^{\lambda_k}$$
$$\frac{d^2}{dt^2} P_{ij}(t) = \sum_{k=1}^{20} c_{ijk} \lambda_k{}^2 e^{\lambda_k} \quad (11)$$

We could do this by direct search, but this would require too much computation burden. We can simplify the problem by adopting the following procedure:

- Input of initial value: From D_{pq}'s, the initial value of θ, denoted by $\hat{\theta}^{(0)}$, is calculated through the least squares that minimizes Eq. (2).
- Renewal of θ: Suppose $\hat{\theta}^{(k-1)} = (t_1^{(k-1)}, \ldots, t_5^{(k-1)})$ is given. Decompose the likelihood function as follows,

$$f(x_1, x_2, x_3, x_4 | \theta) = \sum_{i=1}^{20} \left[\pi_i L^{(1)}(t_1, t_2 | i, x_1, x_2) \times \sum_{j=1}^{20} P_{ij}(t_5) L^{(2)}(t_3, t_4 | j, x_3, x_4) \right] \quad (12)$$

where

$$L^{(1)}(t_1, t_2 | i, x_1, x_2) = P_{ix_1}(t_1) P_{ix_2}(t_2)$$
$$L^{(2)}(t_3, t_4 | j, x_3, x_4) = P_{jx_3}(t_3) P_{jx_4}(t_4)$$

1) Renewal of t_1 and t_2: Calculate $t_1^{(k)}$ and $t_2^{(k)}$ as renewals by Newton's method that maximizes

$$l(t_1, t_2 | t_3^{(k-1)}, t_4^{(k-1)}, t_5^{(k-1)}, \mathbf{X}) = \sum_{h=1}^{n} \log \left[\sum_{i=1}^{20} \pi_i L^{(1)}(t_1, t_2 | i, X_{1h}, X_{2h}) \times \sum_{j=1}^{20} P_{ij}(\hat{t}_5^{(k-1)}) L^{(2)}(\hat{t}_3^{(k-1)}, \hat{t}_4^{(k-1)} | j, X_{3h}, X_{4h}) \right] \quad (13)$$

2) Renewal of t_3 and t_4: Calculate $t_3^{(k)}$ and $t_4^{(k)}$ as renewals by Newton's method that maximizes

$$l(t_3, t_4 | t_1^{(k)}, t_2^{(k)}, t_5^{(k-1)}, \mathbf{X}) = \sum_{h=1}^{n} \log \left[\sum_{i=1}^{20} \pi_i L^{(1)}(\hat{t}_1^{(k)}, \hat{t}_2^{(k)} | i, X_{1h}, X_{2h}) \times \sum_{j=1}^{20} P_{ij}(\hat{t}_5^{(k-1)}) L^{(2)}(t_3, t_4 | j, X_{3h}, X_{4h}) \right] \quad (14)$$

3) Renewal of t_5: Calculate $t_5^{(k)}$ as a renewal by Newton's method that maximizes

$$l(t_5 | t_1^{(k)}, t_2^{(k)}, t_3^{(k)}, t_4^{(k)}, \mathbf{X}) = \sum_{h=1}^{n} \log \left[\sum_{i=1}^{20} \pi_i L^{(1)}(t_1^{(k)}, t_2^{(k)} | i, X_{1h}, X_{2h}) \times \sum_{j=1}^{20} P_{ij}(t_5) L^{(2)}(t_3^{(k)}, t_4^{(k)} | j, X_{3h}, X_{4h}) \right] \quad (15)$$

4) Stopping rule: Stop the procedure when

$$\frac{|t_h^{(k)} - t_h^{(k-1)}|}{t_h^{(k)}} < \epsilon \qquad (\epsilon = 0.01), \qquad h = 1, \ldots, 5$$

otherwise iterate steps from 1 to 3.

A similar procedure is applicable to a tree for more than five OTUs.

Bootstrap Probabilities of Alternative Trees. When we have m candidates for tree topologies (or models), the log-likelihood of each topology is estimated by substituting the above-mentioned maximum likelihood estimates for the parameters. The topology with the highest log-likelihood among m alternatives is chosen as the best candidate for the true tree topology. A likelihood ratio test is generally used in comparing between models. However, it can be applicable only when the models are nested, whereas different bifurcating tree topologies are not nested. Here, we shall estimate the distribution of the likelihood ratio statistic from the variability of log-likelihood among sites, which can be obtained during the process of the maximum likelihood estimation.

The log-likelihoods of m alternative models are represented by

$$l_{(i)}(\theta_{(i)} | \mathbf{X}) = \sum_{h=1}^{n} \log f_{(i)}(\mathbf{X}_h | \theta_{(i)}), \qquad i = 1, \ldots, m \quad (16)$$

where each term of the right-hand side follows an independently identical distribution (i.i.d.). When θ is replaced by the maximum likelihood estimate $\hat{\theta}$, each term of the right-hand side of

$$l_{(i)}(\hat{\theta}_{(i)} | \mathbf{X}) = \sum_{h=1}^{n} \log f_{(i)}(\mathbf{X}_h | \hat{\theta}_{(i)}), \qquad i = 1, \ldots, m \quad (17)$$

no longer follows i.i.d. However, when n is large, the distribution of Eq. (17) coincides asymptotically with that of Eq. (16). Therefore, the estimated log-likelihoods

$$(l_{(1)}, l_{(2)}, \ldots, l_{(m)})$$

asymptotically follow a multivariate normal distribution, whose mean and variance–covariance can be estimated by

$$l_{(i)}(\hat{\theta}_{(i)} | \mathbf{X})$$

and

$$\frac{n}{n-1} \sum_{h=1}^{n} \left[\log f_{(i)}(\mathbf{X}_h | \hat{\theta}_{(i)}) - \frac{1}{n} \sum_{h'=1}^{n} \log f_{(i)}(\mathbf{X}_{h'} | \hat{\theta}_{(i)}) \right] \times \left[\log f_{(j)}(\mathbf{X}_h | \hat{\theta}_{(j)}) - \frac{1}{n} \sum_{h'=1}^{n} \log f_{(j)}(\mathbf{X}_{h'} | \hat{\theta}_{(j)}) \right]$$

(Kishino and Hasegawa 1989), respectively.

It is desirable to carry out bootstrap resampling (Felsenstein 1985; Hasegawa and Kishino 1989) of $\mathbf{X}_h$'s, but it requires too much computational burden. In this paper we shall estimate the bootstrap probability that tree i is selected as the best model from comparison among components of a random number that follows the multivariate normal distribution presented above (an MND method in short). Bootstrap probabilities can be estimated also by bootstrap resampling the estimated log-likelihoods of sites as follows,

$$l_{(i)}^{(p)} = \sum_{h=1}^{n} \log f_{(i)}(\mathbf{X}_h^{(p)} | \hat{\theta}_{(i)}) \quad (18)$$

where $\mathbf{X}_h^{(B)}$ refers to a site resampled by bootstrap (a resampling estimated log-likelihood method, or RELL method) (Felsenstein, personal communication). Both methods can give bootstrap probabilities for candidate trees without performing a maximum likelihood estimation for each resampled data set.

In the Case of Many OTUs. Because the maximum likelihood method presented in this paper explosively consumes a large amount of CPU time as the number of OTUs increases, it might be impractical to employ the method for more than 6 OTUs. Therefore, we develop here a simplified method for inferring trees among groups of many OTUs.

Suppose that we have $s > 5$ OTUs, and that they are known to be clustered in advance into five groups that contain s_1, s_2, s_3, s_4, and s_5 members, respectively ($s = s_1 + s_2 + s_3 + s_4 + s_5$). Then, sequence data can be represented as follows:

$$\text{Group 1}\begin{cases}\text{Species (1.1)} & X_{(1.1)1} & X_{(1.1)2} & \cdots & X_{(1.1)n}\\ \vdots & & \vdots & & \\ \text{Species (1.}s_1\text{)} & X_{(1.s_1)1} & X_{(1.s_1)2} & \cdots & X_{(1.s_1)n}\end{cases}$$

$$\vdots$$

$$\text{Group 5}\begin{cases}\text{Species (5.1)} & X_{(5.1)1} & X_{(5.1)2} & \cdots & X_{(5.1)n}\\ \vdots & & \vdots & & \\ \text{Species (5.}s_5\text{)} & X_{(5.s_5)1} & X_{(5.s_5)2} & \cdots & X_{(5.s_5)n}\end{cases}$$

Picking up one species from each group, we apply the maximum likelihood procedure described above for m alternative tree topologies of the five groups. If we examine all possible tree topologies for 5 OTUs, m is 15. Then, we get $r = s_1 \times s_2 \times s_3 \times s_4 \times s_5$ sets of estimates of branch lengths and log-likelihoods for each model i ($i = 1, \ldots, n$), i.e.,

$$\hat{\theta}^j_{(i)} \text{ and } l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j), \quad j = 1, 2, \ldots, r$$

where $\mathbf{X}^j$ is the $5 \times n$ submatrix of the $r \times n$ full data, corresponding to the selected five species. Variances and covariances among $l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j)$ ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, r$) are estimated by

$$\widehat{\operatorname{cov}}\left[l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j),\ l_{(i')}(\hat{\theta}^{j'}_{(i')}|\mathbf{X}^{j'})\right] = \frac{n}{n-1}\sum_{h=1}^{n}\left[\log f_{(i)}(\mathbf{X}^j_h|\hat{\theta}^j_{(i)}) - \frac{1}{n}\sum_{h'=1}^{n}\log f_{(i)}(\mathbf{X}^j_{h'}|\hat{\theta}^j_{(i)})\right] \times\left[\log f_{(i')}(\mathbf{X}^{j'}_h|\hat{\theta}^{j'}_{(i')}) - \frac{1}{n}\sum_{h'=1}^{n}\log f_{(i')}(\mathbf{X}^{j'}_{h'}|\hat{\theta}^{j'}_{(i')})\right] \quad (19)$$

(Kishino and Hasegawa 1989). Sets of r estimated log-likelihoods are averaged with weights inversely proportional to the variances:

$$l_{(i)} = \sum_{j=1}^{r} w^j_{(i)} l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j) \quad (20)$$

where

$$w^j_{(i)} = \frac{1/\widehat{\operatorname{var}}\left[l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j)\right]}{\sum_{j'=1}^{r} 1/\widehat{\operatorname{var}}\left[l_{(i)}(\hat{\theta}^{j'}_{(i)}|\mathbf{X}^{j'})\right]} \quad (21)$$

Variances and covariances among them are given by

$$\widehat{\operatorname{cov}}\left[l_{(i)}, l_{(i')}\right] = \sum_{j=1}^{r}\sum_{j'=1}^{r} w^j_{(i)} w^{j'}_{(i')} \widehat{\operatorname{cov}}\left[l_{(i)}(\hat{\theta}^j_{(i)}|\mathbf{X}^j),\ l_{(i')}(\hat{\theta}^{j'}_{(i')}|\mathbf{X}^{j'})\right] \quad (22)$$

Estimation of Parameters. Furthermore, the parameters $\theta_{(i)}$ ($i = 1, 2, \ldots, m$) are estimated by

$$\hat{\theta}_{(i)} = \sum_{j=1}^{r} w^j_{(i)} \hat{\theta}^j_{(i)} \quad (23)$$

Their variances are given by

$$\operatorname{var}(\hat{\theta}_{(i)}) = \sum_{j=1}^{r}\sum_{j'=1}^{r} w^j_{(i)} w^{j'}_{(i)} \operatorname{cov}(\hat{\theta}^j_{(i)}, \hat{\theta}^{j'}_{(i)}) \quad (24)$$

The covariances in the right-hand side are given by

$$n \operatorname{cov}(\hat{\theta}^j_{(i)}, \hat{\theta}^{j'}_{(i)}) = (J^j_{(i)})^{-1} K^{jj'}_{(i)} (J^{j'}_{(i)})^{-1} \quad (25)$$

where

$$J^j_{(i)} = -E\left[\nabla\nabla^T \log f_{(i)}(\mathbf{X}^j|\theta^j_{(i)})\right] \quad (26)$$

and

$$K^{jj'}_{(i)} = E\left[\nabla \log f_{(i)}(\mathbf{X}^j|\theta^j_{(i)})\nabla^T \log f_{(i)}(\mathbf{X}^{j'}|\theta^{j'}_{(i)})\right] \quad (27)$$

which are estimated by

$$-\frac{1}{n}\sum_{h=1}^{n}\nabla\nabla^T \log f_{(i)}(\mathbf{X}^j_h|\hat{\theta}^j_{(i)})$$

and

$$\frac{1}{n}\sum_{h=1}^{n}\nabla \log f_{(i)}(\mathbf{X}^j_h|\hat{\theta}^j_{(i)})\nabla^T \log f_{(i)}(\mathbf{X}^{j'}_h|\hat{\theta}^{j'}_{(i)})$$

respectively.

Markov Model of Insertion/Deletion Events. Up to now in this paper, we have been concerned only with amino acid substitutions and have not taken into account insertion/deletion events in amino acid sequences. However, as the rate of the latter events is generally lower than that of the former, they, if they occurred, should provide important information for inferring evolutionary trees. There have been several attempts to infer trees taking into account insertions and deletions in addition to substitutions (Meyer et al. 1986; Hasegawa et al. 1987; Morden and Golden 1989a). However, because of the difficulty in modeling the insertion/deletion events compared to substitutions, few quantitative attempts have been made. DNA fingerprinting (Jeffreys et al. 1985; Lynch 1988), which has become popular in forensic science and also in evolutionary sociobiology, uses information about insertion/deletion events that occur in extremely high frequency in tandemly repeating sequences for estimating genetic relatedness between individuals. Contrary to that situation, we assume that the frequency of insertion/deletion events is low enough that the sequence length does not change significantly during the time scale under consideration.

As a first approximation, we consider a model where insertion/deletion occurs independently among sites. This can be formulated by a Markov process of transition between + (presence of a stretch of sequences) and − (its absence) states. Once a long stretch of sequences is deleted, it is almost impossible to recover the same stretch of sequences except in some cases such as in the case of tandem repetition. However, we do not look at the exact order of amino acids in a stretch of sequences when modeling the insertion/deletion, and just look at the presence or absence of a stretch of sequences. Let v be the rate of insertions/deletions per site per unit time; that is, both the rates of insertion and deletion of a stretch of sequences are assumed to be $v/2$. Then the transition probability during an infinitesimally short time interval, dt, is represented by

$$P(dt) = \begin{array}{c} \\ + \\ - \end{array}\begin{array}{c} \begin{array}{cc} + & - \end{array} \\ \begin{bmatrix} 1 - v\,dt/2 & v\,dt/2 \\ v\,dt/2 & 1 - v\,dt/2 \end{bmatrix} \end{array} \quad (28)$$

The transition probability matrix for an arbitrary time interval t is given by

$$P(t) = \frac{1}{2}\begin{bmatrix} 1 + e^{-w} & 1 - e^{-w} \\ 1 - e^{-w} & 1 + e^{-w} \end{bmatrix} \quad (29)$$

A time interval during which one amino acid substitution occurs per 100 amino acids is taken as a unit of time as before. Similarly, as in Eq. (9), a probability of occurrence of a particular pattern of distribution of + and − states among species in a site can be calculated. Let $q(\theta, v)$ denote the sum of the probabilities of $(+, +, \ldots, +)$ and of $(-, -, \ldots, -)$.

Letting ν denote the number of stretches of sequences that have information on insertions/deletions, we represent the pattern of insertions/deletions by $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_\nu)$, where $\mathbf{Y}_i$ is a column vector of dimension s and its elements are either + or −, and their likelihoods are denoted by $g(\mathbf{Y}_1|\theta, v)$, $g(\mathbf{Y}_2|\theta, v)$, $\ldots$, $g(\mathbf{Y}_\nu|\theta, v)$. Denoting the likelihoods of respective sites for the amino acid substitutions derived before by $f(\mathbf{X}_1|\theta)$, $f(\mathbf{X}_2|\theta)$, $\ldots$, $f(\mathbf{X}_n|\theta)$, the total likelihood of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_\nu)$ that takes into account insertion/deletion events is given by

$$L(\theta, v|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_\nu) = \prod_{i=1}^{n} q(\theta, v) f(\mathbf{X}_i|\theta) \prod_{j=1}^{\nu} g(\mathbf{Y}_j|\theta, v) \quad (30)$$

The log-likelihood is

$$l = l_s(\theta|\mathbf{X}) + l_{ID}(\theta, v|\mathbf{Y}) \quad (31)$$

where

$$l_s(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log f(\mathbf{X}_i|\theta) \quad (32)$$

and

$$l_{ID}(\theta, v|\mathbf{Y}) = n \log q(\theta, v) + \sum_{j=1}^{\nu} \log g(\mathbf{Y}_j|\theta, v) \quad (33)$$

Equation (32) gives information on amino acid substitutions and Eq. (33) on insertion/deletion events.

Although the latter should contribute in the selection of tree topologies, it is expected that it does not contribute significantly toward estimating the branch lengths, θ. Therefore our procedure is as follows: (1) at first, we estimate $\hat{\theta}$ from the partial log-likelihood l_s, and (2) using $\hat{\theta}$, we estimate $\hat{v}$ by maximizing $l_{ID}(\hat{\theta}, v|\mathbf{Y})$. The variance of the estimated v is given by

$$\widehat{\mathrm{var}}(\hat{v}) = \left[\frac{\partial^2}{\partial v^2} l_{ID}(\hat{\theta}, \hat{v}|\mathbf{Y})\right]^{-1} + \left[\frac{\partial}{\partial v}\nabla_\theta^T l_{ID}(\hat{\theta}, \hat{v}|\mathbf{Y})\right][\nabla_\theta \nabla_\theta^T l_s(\hat{\theta}|\mathbf{X})]^{-1}\nabla_\theta \frac{\partial}{\partial v} l_{ID}(\hat{\theta}, \hat{v}|\mathbf{Y}) \quad (34)$$

The first term represents the variance given the estimates of branch lengths, and the second term represents the effect of the variance of the estimates of branch lengths.

Bootstrap probabilities can be estimated in the same way as before by representing the log-likelihood of $\tilde{\mathbf{X}}_i$ $(i = 1, 2, \ldots, n + \nu)$ by

$$\log \tilde{f}(\tilde{\mathbf{X}}_i|\theta, v) = \begin{cases} \log q(\theta, v) + \log f(\mathbf{X}_i|\theta) & (i = 1, 2, \ldots, n) \\ \log g(\mathbf{Y}_{i-n}|\theta, v) & (i = n + 1, n + 2, \ldots, n + \nu) \end{cases} \quad (35)$$

Phylogeny of *psb*A and the Origin of Chloroplasts

Recently, two groups reported on the phylogenetic position of *Prochlorothrix*, a presumed relative of *Prochloron*, in relation to the origin of chloroplasts of green plants, but reached apparently conflicting conclusions (Morden and Golden 1989a; Turner et al. 1989; for a review see Penny 1989). Turner et al. (1989) used nucleotide sequence data of 16S rRNA and placed *Prochlorothrix* apart from green chloroplasts. On the other hand, Morden and Golden (1989a) used amino acid sequence data of a photosynthetic membrane protein, *psb*A from photosystem II, and reached the opposite conclusion, indicating a common ancestry for *Prochlorothrix* and green chloroplasts. As a confidence limit was not attached to Morden and Golden's tree, the apparent contradiction between the two groups may not be real. As Penny (1989) says, any measurement, to be scientific, must include an indication of its accuracy. Trees without indication of their accuracy will cause much confusion.

The method developed in this paper can provide confidence limits for the inferred tree. This method is not sensitive to unequal rates of evolution among different lineages, as is apparently the case in the *psb*A tree, whereas the parsimony method used by Morden and Golden is sensitive (Felsenstein 1978; Hasegawa and Yano 1984). It has been shown recently that parsimony cannot guarantee a correct result even with equal rates for different lineages (Penny et al. 1987; Hendy and Penny 1989).

We applied our method to the amino acid sequence data of *psb*A. The maximum likelihood analysis was carried out for five *psb*A sequences; that is, from *Fremyella displosiphon* (Mulligan et al. 1984), *Synechocystis* 6803 (Osiewacz and McIntosh 1987), *Anacystis nidulans* R2 (also called *Synechococcus* sp. strain 7942; *psb*AII) (Golden et al. 1986), *Prochlorothrix hollandica* (Morden and Golden 1989a), and a green chloroplast. As a representative of green chloroplasts, one among the four species, that is, *Nicotiana tabacum* (Shinozaki et al. 1986), *Petunia hybrida* (Aldrich et al. 1986), *Marchantia polymorpha* (Ohyama et al. 1986), and *Chlamydomonas reinhardii* (Erickson et al. 1984), was chosen, and the estimates were averaged over the four choices by the method described in the preceding section. A tree inferred by our method is unrooted, and 15 unrooted trees are possible for 5 OTUs. All possible alternatives were examined.

The result is shown in Table 1. At first, we shall be concerned only with amino acid substitutions (without insertion/deletion). The tree with the highest log-likelihood is tree 1, where *Prochlorothrix* links with *Anacystis* and chloroplasts link with *Fremyella*, and the estimated bootstrap probability of tree 1, P_1, is 0.415 (by the MND method). To the contrary, tree 9 that is suggested by Morden and Golden (1989a) has a log-likelihood lower by 11.08 ± 10.88 (± 1 SE) than tree 1, and P_9 is only 0.079. The sub-

Table 1. Comparison among 15 alternative unrooted trees of *psb*A by the maximum likelihood method

Tree topology		Without insertion/deletion		With insertion/deletion		
		$l_i - l_1$	P_i	$l_i - l_1$	P_i	$\hat{v}_i$
1	(A.n, chl) | (Pro, Syn, Fre)	0	0.415 (0.416)	0	0.377 (0.365)	0.0280 ±0.0201
2	(Fre, chl) | (Pro, Syn, A.n)	−19.21 ± 11.00	0.000 (0.000)	−19.08 ± 11.01	0.000 (0.000)	0.0268 ±0.0192
3	(Fre, chl) | (A.n, Syn, Pro)	−16.19 ± 11.45	0.003 (0.002)	−10.74 ± 12.29	0.009 (0.008)	0.0138 ±0.0140
4	(A.n, chl) | (Pro, Fre, Syn)	−1.68 ± 5.19	0.273 (0.260)	−1.69 ± 5.19	0.218 (0.224)	0.0281 ±0.0201
5	(Syn, chl) | (Pro, Fre, A.n)	−15.30 ± 11.77	0.021 (0.018)	−15.34 ± 11.77	0.012 (0.016)	0.0273 ±0.0196
6	(Syn, chl) | (A.n, Fre, Pro)	−15.98 ± 11.47	0.006 (0.006)	−10.55 ± 12.30	0.016 (0.015)	0.0135 ±0.0136
7	(Fre, chl) | (Pro, A.n, Syn)	−16.12 ± 10.13	0.004 (0.003)	−16.14 ± 10.13	0.002 (0.002)	0.0272 ±0.0195
8	(Syn, chl) | (Pro, A.n, Fre)	−7.57 ± 8.44	0.119 (0.121)	−7.77 ± 8.44	0.088 (0.095)	0.0275 ±0.0197
9	(Syn, chl) | (Fre, A.n, Pro)	−11.08 ± 10.88	0.079 (0.082)	−5.63 ± 11.76	0.219 (0.208)	0.0138 ±0.0139
10	(Fre, chl) | (A.n, Pro, Syn)	−16.98 ± 9.78	0.001 (0.000)	−16.96 ± 9.78	0.000 (0.000)	0.0271 ±0.0194
11	(Syn, chl) | (A.n, Pro, Fre)	−12.40 ± 6.88	0.001 (0.001)	−12.38 ± 6.88	0.001 (0.001)	0.0266 ±0.0191
12	(Syn, chl) | (Fre, Pro, A.n)	−14.47 ± 10.27	0.008 (0.013)	−14.32 ± 10.27	0.004 (0.008)	0.0274 ±0.0197
13	(Fre, Pro) | (A.n, chl, Syn)	−17.54 ± 11.08	0.000 (0.000)	−17.50 ± 11.08	0.000 (0.000)	0.0270 ±0.0194
14	(Syn, Pro) | (A.n, chl, Fre)	−22.08 ± 10.09	0.000 (0.000)	−21.91 ± 10.09	0.000 (0.000)	0.0263 ±0.0189
15	(Syn, Pro) | (Fre, chl, A.n)	−3.33 ± 4.58	0.071 (0.078)	−3.25 ± 4.58	0.053 (0.059)	0.0278 ±0.0200

Fre = *Fremyella displosiphon*, Syn = *Synechocystis*, A.n = *Anacystis nidulans*, Pro = *Prochlorothrix hollandica*, chl = chloroplasts. Variances (± denotes 1 SE) of log-likelihood difference and of $\hat{v}$ were calculated by Eq. (12) in Kishino and Hasegawa (1989) and by Eq. (34) in the text, respectively. Bootstrap probability, P_i, of tree i being the maximum likelihood tree among 15 alternatives was estimated by the MND method and also by the RELL method (shown in parentheses) (sample size of 10^4)

total of bootstrap probabilities of trees that link *Prochlorothrix* with chloroplasts, $P_3 + P_6 + P_9$, is 0.087, and such clustering seems unlikely by this analysis. The subtotal of bootstrap probabilities of trees that link *Prochlorothrix* with *Anacystis*, $P_1 + P_4 + P_{15}$, is 0.759, and such clustering seems likely.

Morden and Golden (1989a) suggested *Prochlorothrix*/chloroplast clustering initially from the par-

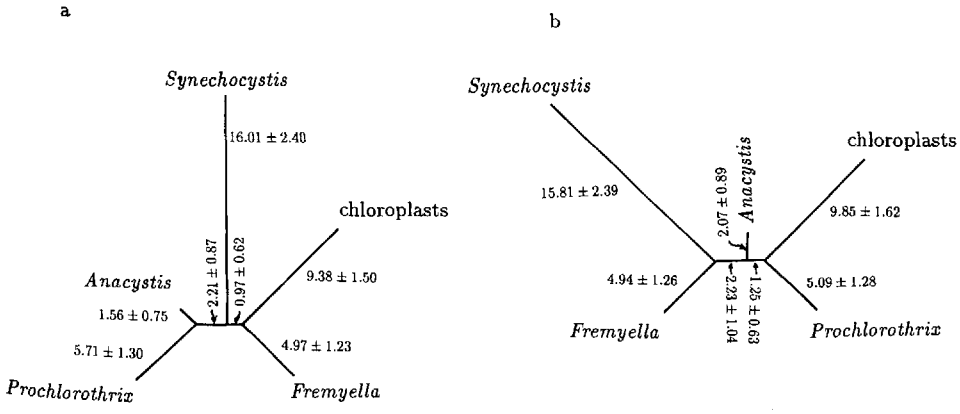

Fig. 2. The maximum likelihood tree (a: tree 1) and the tree proposed by Morden and Golden (b: tree 9) of *psb*A. Length of a branch is proportional to the estimated number of amino acid substitutions per 100 amino acids that is shown with its SE (±).

simony analysis of amino acid substitutions as well as from the analysis of insertion/deletion data, but their parsimony analysis has turned out to be in error (Morden and Golden 1989b). The *Prochlorothrix*/chloroplast clustering (tree 9) is only equally likely with tree 1 from the parsimony analysis of amino acid substitutions. Now, the sole molecular data that seem to support their tree are the insertion/deletion data of *psb*A, i.e., deletion of a stretch of seven amino acids near the C terminus shared by *Prochlorothrix* and chloroplasts (Fig. 2 in Morden and Golden 1989a). The significance of these data also is evaluated in Table 1.

When insertion/deletion is taken into account (with insertion/deletion in Table 1), P_9 is 0.219 and P_1 is 0.377. The subtotal of bootstrap probabilities of trees that link *Prochlorothrix* with chloroplasts, $P_3 + P_6 + P_9$, is 0.244, and such clustering cannot be rejected. The subtotal of bootstrap probabilities of trees that link *Prochlorothrix* with *Anacystis*, $P_1 + P_4 + P_{15}$, is still as high as 0.648. As *Fremyella* is lacking in the 16S rRNA tree of Turner et al. (1989), all of trees 1, 4, and 15 are consistent with their tree, and the high bootstrap probability of this link is in accord with their analysis. Although the insertion/deletion data favor the *Prochlorothrix*/chloroplast clustering as was claimed by Morden and Golden, they are not strong enough to refute the conclusion obtained by the amino acid substitution data, and the apparent contradiction between the two research groups seems to be an artifact caused by defects of the methods used for data analysis.

The maximum likelihood estimates of v are shown in the last column of Table 1. These represent rates of insertions/deletions relative to that of amino acid substitutions. The estimates of v for trees 3, 6, and 9 are nearly half of those for other trees. This is due to the fact that, in the trees other than trees 3, 6, and 9, parallel deletions along the chloroplasts and *Prochlorothrix* lines must be assumed. In Fig. 2, tree 1 (the maximum likelihood tree) and tree 9 (Morden and Golden's tree) are shown with the estimated branch lengths. It should be noted that in tree 9 the branch length of the common ancestral line between the chloroplasts and *Prochlorothrix* after separating from the others is only 1.25 ± 0.63 substitutions per 100 amino acids, and that the deletion common to these two taxa, if they are shared derived characters as Morden and Golden claim, should have occurred along this short branch. By contrast, in tree 1, although two parallel deletions must have occurred along the lines leading to *Prochlorothrix* and to chloroplasts, these lines may be long (5.71 ± 1.30 and 9.38 ± 1.50, respectively) enough to allow such parallel deletions. For these reasons the deletion data are not as strong as was claimed by Morden and Golden. It must be noted that we have assumed homogeneity of the insertion/deletion probability among sites. If some stretches of sequences are apt to delete or insert because of tandem repetition (Hasegawa et al. 1987) or for other unknown reasons, the strength of insertion/deletion data may be further weakened.

In Table 1, bootstrap probabilities estimated by the RELL method are shown in parentheses as well as those estimated by the MND method. It is apparent that these two methods give essentially identical estimates. Kishino and Hasegawa (unpublished) have shown that the two methods give good approximations of bootstrap probabilities obtained by performing maximum likelihood estimation for bootstrap resampled data sets. The RELL method is simple in procedure, whereas the MND method may be useful when the sequence length is large because its CPU time does not depend on the length of the data set.

Discussion

One may wonder whether Dayhoff et al.'s transition matrix is applicable to the evolution of *psb*A. The matrix was constructed on the basis of the sequence data for many different proteins among which hemoglobins and cytochrome c together occupy a major portion. However, we do not think that the

matrix is specific only for these proteins, because the frequency of substitutions in the matrix correlates well with the chemical similarity between amino acids (Clarke 1970). This indicates that amino acid substitutions that are accompanied by small chemical changes occur much more frequently in evolution than those accompanied by large ones, which is consistent with the neutral theory (Kimura 1983). Therefore, it seems reasonable to assume that the evolution of *psb*A follows Dayhoff's matrix if most of the evolutionary changes of this protein are neutral, as is likely to be the case.

Furthermore, some readers may question our assumption of the identical transition probability matrix among different sites. It is known that the variability depends not only on the amino acid species but on the site. Actually, however, there is a high correlation between site dependency and amino acid dependency, and therefore the site dependency of variability is taken into account to some extent in Dayhoff et al.'s transition matrix (Nei and Tateno 1978).

We have specified a very simple model for insertion/deletion events, because the mechanism is not well known at present. We recognize the immaturity of the model. More elaborate considerations are left for future study.

It might be interesting to point out that even with insertion/deletion the log-likelihood of tree 9 is lower than that of tree 4 and tree 15 ($l_9 - l_1 = -5.63$, $l_4 - l_1 = -1.69$, $l_{15} - l_1 = -3.25$), whereas P_9 is higher than P_4 and P_{15}. This is due to an extremely large variance of $l_9 - l_1$ (more than five times of those of the other two). Therefore, it is apparent that the mean log-likelihood differences among models cannot give sufficient information on the reliability of the maximum likelihood model unless the variance of the difference is attached (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989).

Prochlorothrix, like *Prochloron,* differs from cyanobacteria by the lack of phycobilin pigments and the possession of chlorophyll *b* in addition to chlorophyll *a,* a combination found in chloroplasts (Burger-Wiersma 1986). For this reason, although there is no direct evidence from proteins or nucleic acids to link *Prochlorothrix* with *Prochloron,* some biologists assume that these two organisms form a clade called Prochlorophyta and that this group is the most closely related to chloroplasts (Miller and Jacob 1989). Although the hypothesis of a Prochlorophyta/chloroplast linkage cannot be excluded when the deletion in *psb*A is taken into account, our analysis indicates that the most likely hypothesis is that *Prochlorothrix* is more closely related to *Anacystis* rather than to the chloroplasts, which is consistent with Turner et al. (1989). Therefore, the resemblance of prochlorophytes to green chloroplasts may be a result of convergent evolution (Cavalier-Smith 1982; Turner et al. 1989). In the 16S rRNA tree, *Cyanophora paradoxa* is the closest relative to the chloroplasts. Unfortunately, the sequence data of *psb*A from this organism are still unpublished. The data should shed more light on the origin of the chloroplasts.

In recent years it has become popular to use bootstrapping in molecular phylogenetics. This tendency is a great advancement in this field, because several years ago it was rare that an attempt was made to assign a confidence interval to an estimated phylogeny. However, the following warning by Felsenstein (1985) who has been advocating the use of bootstrapping in phylogenetics should be noted.

> Bootstrapping provides us with a confidence interval within which is contained not the true phylogeny, but the phylogeny that would be estimated on repeated sampling of many characters from the underlying pool of characters. As such it may be misleading if the method used to infer phylogenies is inconsistent.

Therefore, even if bootstrapping excludes alternative trees by a method that is inconsistent in a particular situation, the inferred tree is not necessarily the truth.

The parsimony method has some merits because it is easy to interpret and also to calculate. However, it sometimes gives erroneous results, particularly when the evolutionary rate differs among lineages (Felsenstein 1978; Hasegawa and Yano 1984; Hendy and Penny 1989). The maximum likelihood method requires a statistical model, and usually this requires more computation than the other approaches. However, one of the greatest advantages of the maximum likelihood method is that it can indicate the accuracy of the inferred tree as we have seen in this paper. Therefore, we will not draw an incorrect conclusion even if the highest likelihood tree is not the correct tree due to random fluctuations in evolution. In that case we simply conclude that the data cannot discriminate among alternative trees. In the maximum likelihood framework, the expected posterior probability of a tree topology is approximately the same as the bootstrap probability (unpublished).

Because the maximum likelihood method developed in this paper takes into account the unequal transition probabilities among pairs of amino acids based on the empirical data of Dayhoff et al. (1978), and because it does not assume constancy of the evolutionary rate among lineages, it appears to be applicable to a wide range of phylogenetic problems. Furthermore, as we have seen in the analysis of insertion/deletion data, the maximum likelihood method can take into account various kinds of information on the same basis. Therefore, we can im-

prove the model as further information accumulates. The problem of computational burden should improve in the near future as computer facilities are rapidly developing and this approach is thus expected to become important in molecular phylogenetics.

Acknowledgments. In the analysis of insertion/deletion data, we used the numerical optimization program UCOP1 developed by Miss S. Ueda at the Institute of Statistical Mathematics. We are grateful to her for allowing us to use the program. We thank Mr. M. Abe, Mr. J. Adachi, and Miss M. Fujiwara for computer programming. We also thank Dr. F. Tajima and an anonymous reviewer for their valuable comments on an earlier version of the manuscript. This study was supported by grants from the Ministry of Education, Science, and Culture of Japan.

References

Aldrich J, Cherney B, Merlin E (1986) Sequence of the chloroplast-encoded *psb*A gene for the Q_B polypeptide of alfalfa. Nucleic Acids Res 14:9537

Bishop MJ, Friday AE (1985) Evolutionary trees from nucleic acid and protein sequences. Proc R Soc Lond B 226:271–302

Bishop MJ, Friday AE (1987) Tetrapod relationships: the molecular evidence. In: Patterson C (ed) Molecules and morphology in evolution: conflict or compromise? Cambridge University Press, Cambridge, pp 123–139

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Burger-Wiersma T, Veenhuis M, Korthals HJ, Van de Wiel CCM, Mur LR (1986) A new prokaryote containing chlorophylls a and b. Nature 320:262–264

Cavalier-Smith T (1982) The origins of plastids. Biol J Linn Soc 17:289–306

Chakraborty R (1977) Estimation of time of divergence from phylogenetic studies. Can J Genet Cytol 19:217–223

Clarke B (1970) Selective constraints on amino-acid substitutions during the evolution of proteins. Nature 228:159–160

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

Erickson JM, Rahire M, Rochaix J-D (1984) *Chlamydomonas reinhardii* gene for the 32,000 mol. wt. protein of photosystem II contains four large introns and is located entirely within the chloroplast inverted repeat. EMBO J 3:2753–2762

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

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

Felsenstein J (1983a) Methods for inferring phylogenies: a statistical view. In: Felsenstein J (ed) Numerical taxonomy. Springer-Verlag, Berlin, pp 315–334

Felsenstein J (1983b) Statistical inference of phylogenies. J R Stat Soc A 146:246–272

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Golden SS, Brusslan J, Haselkorn R (1986) Expression of a family of *psb*A genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* r2. EMBO J 5:2789–2798

Hasegawa M, Kishino H (1989) Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution 43:672–677

Hasegawa M, Yano T (1984) Maximum likelihood method of phylogenetic inference from DNA sequence data. Bull Biomet Soc Jpn 5:1–7

Hasegawa M, Iida Y, Yano T, Takaiwa F, Iwabuchi M (1985) Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. J Mol Evol 22:32–38

Hasegawa M, Kishino H, Yano T (1987) Man's place in Hominoidea as inferred from molecular clocks of DNA. J Mol Evol 26:132–147

Hasegawa M, Kishino H, Yano T (1988) Phylogenetic inference from DNA sequence data. In: Matusita K (ed) Statistical theory and data analysis II: Proceedings of Second Pacific Area Statistical Conference. North-Holland, Amsterdam, pp 1–13

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

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Kikuno R, Hayashida H, Miyata T (1985) Rapid rate of rodent evolution. Proc Jpn Acad B 61:153–156

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kishino H, Hasegawa M (1989) Evolution 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

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599

Meyer TE, Cusanovich MA, Kamen MD (1986) Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. Proc Natl Acad Sci USA 83: 217–220

Miller KR, Jacob JS (1989) On *Prochlorothrix*. Nature 338: 303–304

Morden CW, Golden SS (1989a) *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts. Nature 337: 382–385

Morden CW, Golden SS (1989b) *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts (Corrigendum). Nature 339:400

Mulligan B, Schultes N, Chen L, Bogorad L (1984) Nucleotide sequence of a multiple-copy gene for the B protein of photosystem II of a cyanobacterium. Proc Natl Acad Sci USA 81:2693–2697

Nei M, Tateno Y (1978) Nonrandom amino acid substitution and estimation of the number of nucleotide substitutions in evolution. J Mol Evol 11:333–347

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572–574

Osiewacz HD, McIntosh L (1987) Nucleotide sequence of a member of the *psb*A multigene family from the unicellular cyanobacterium *Synechocystis* 6803. Nucleic Acids Res 15: 10585

Penny D (1989) What, if anything, is *Prochloron*? Nature 337: 304–305

Penny D, Hendy MD, Henderson IM (1987) Reliability of evolutionary trees. Cold Spring Harbor Symp Quant Biol 52:857–862

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Sinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The

complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Turner S, Burger-Wiersma T, Giovannoni SJ, Mur LR, Pace NR (1989) The relationship of a prochlorophyte *Prochlorothrix hollandica* to green chloroplasts. Nature 337:380–382

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Received October 17, 1989/Revised December 25, 1989