A Covariotide Model Explains Apparent Phylogenetic Structure of Oxygenic Photosynthetic Lineages

P. J. Lockhart, M. A. Steel,† A. C. Barbrook,‡ D. H. Huson,§ M. A. Charleston,‖ and C. J. Howe‡*

*Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand; †Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand; ‡Department of Biochemistry, University of Cambridge, UK; §Program in Applied and Computational Mathematics, Princeton University; and ‖Department of Zoology, University of Oxford

The aims of the work were (1) to develop statistical tests to identify whether substitution takes place under a covariotide model in sequences used for phylogenetic inference and (2) to determine the influence of covariotide substitution on phylogenetic trees inferred for photosynthetic and other organisms. (Covariotide and covarion models are ones in which sites that are variable in some parts of the underlying tree are invariable in others and vice versa.) Two tests were developed. The first was a contingency test, and the second was an inequality test comparing the expected number of variable sites in two groups with the observed number. Application of these tests to 16S rDNA and *tuf*A sequences from a range of nonphotosynthetic prokaryotes and oxygenic photosynthetic prokaryotes and eukaryotes suggests the occurrence of a covariotide mechanism. The degree of support for partitioning of taxa in reconstructed trees involving these organisms was determined in the presence or absence of sites showing particular substitution patterns. This analysis showed that the support for splits between (1) photosynthetic eukaryotes and prokaryotes and (2) photosynthetic and nonphotosynthetic organisms could be accounted for by patterns arising from covariotide substitution. We show that the additional problem of compositional bias in sequence data needs to be considered in the context of patterns of covariotide/covarion substitution. We argue that while covariotide or covarion substitution may give rise to phylogenetically informative patterns in sequence data, this may not always be so.

Introduction

Sequence data have been used extensively in studying the development of oxygenic photosynthesis and the subsequent endosymbiotic origins of plastids in photosynthetic eukaryotes. A question of particular interest has been whether plastids with different light-harvesting pigment types arose independently or from a single endosymbiosis (monophyletically). Many studies have concluded that plastids had a monophyletic origin (e.g., Palmer 1993; Delwiche, Kuhsel, and Palmer 1995). However, we observed in an earlier paper (Lockhart et al. 1992) that base composition bias could contribute significantly to apparent phylogenetic structure among sequences from plastids and oxygenic photosynthetic bacteria (chloroxybacteria) and give rise to support for a monophyletic origin that might not necessarily reflect a genuine historical relationship. The importance of compositional bias in phylogenetic reconstruction is now widely recognized (e.g., Hasegawa and Hashimoto 1993; Lake 1994; Lockhart et al. 1994; Pesole et al. 1995; Jermiin et al. 1996). In considering the significance of compositional bias, we emphasized the importance of understanding the distribution of sites that are free to vary in a given alignment (Lockhart et al. 1992). Here, we characterize the complexity of this issue and show that a covariotide pattern of substitution describes the evolution of oxygenic photosynthetic lineages.

Assumptions about the distribution of variable/invariable sites are important in evolutionary tree reconstruction. It is useful to distinguish three different model types: type 1—Markov models under which there is no variation in rates of change between different sequence positions; type 2—‘‘rates-across-sites’’ (RAS) Markov models under which different sequence positions can change at different rates (It is useful to subdivide this class of models further into type 2.1—models in which a certain subset of sites are invariable [i.e., evolving at rate 0] but the remaining sites all evolve at a constant rate, and type 2.2—models in which there may be invariable sites, and the variable sites are evolving at different rates.); and type 3—covariotide/covarion models under which sites that are invariable in one part of the underlying tree can be variable in another and vice versa. (As defined by Shoemaker and Fitch [1989], ‘‘covariotides’’ refer to nucleotide sequences, while ‘‘covarions’’ refer to protein sequences).

Some covariotide and covarion patterns of change can mislead evolutionary tree building. This occurs when distantly related sequences share more similar distributions of invariable sites than do closely related species (Lockhart et al. 1996). Consequently, patterns arise under an extreme form of the model first described by Felsenstein (1978) which can lead to inconsistency. This problem is undetected even by maximum-likelihood tree reconstruction methods if invariable sites are not recognized as being present in the data and if some of the invariable sites also occur at the same sequence positions in all taxa (Lockhart et al. 1996).

Here, we describe a contingency test which can reject some noncovariotide/noncovarion models (specifically, types 1 and 2.1) and show that this occurs for sequences used in phylogenetic reconstruction for pho-

Abbreviations: RAS, rates across sites; HKY model, Hasegawa-Kishino-Yano (1985) substitution model.

Key words: phylogenetic inference, covarion, covariotide, compositional bias, plastid origins, evolution of oxygenic photosynthesis.

Address for correspondence and reprints: P. J. Lockhart, Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand. E-mail: p.j.lockhart@massey.ac.nz.

Mol. Biol. Evol. 15(9):1183–1188. 1998

tosynthetic organisms. We also describe a second inequality test for pairwise comparison of groups of sequences to test if substitution follows a covariotide or covarion model. The test indicates that this may indeed be the case for 16S and *tuf*A sequences. We discuss the observed phylogenetic structures of sequences from photosynthetic organisms in light of the substitution patterns indicated.

Materials and Methods

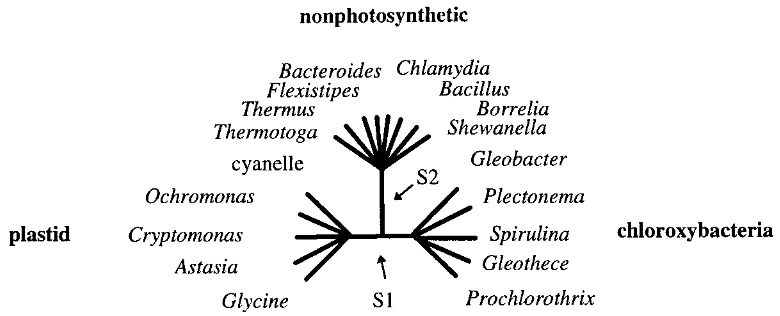
Aligned 16S rDNA sequences were extracted from the RDP database (Olsen, Woese and Overbeek 1994; http://rdp.life.uiuc.edu/). Aligned eubacterial *tuf*A sequences were taken from Delwiche, Kuhsel, and Palmer (1995). RDP loci and GenBank accession numbers, where available for taxa, are given repectively for 16S rDNA and *tuf*A sequences. These were for chloroxybacteria (*Prochlorothrix hollandica* [Prtx.holla; U09445], *Gleobacter violaceus* [Glb.violac; U09433], *Gleothece* sp. [Glth.membr; U09434], *Spirulina* sp. [Spli.sp2; X75044, X15646], *Plectonema boryanum* [Plec.borya; U09444], plastids (*Astasia longa* [Asta.lon_C; X14386, X14385], *Cryptomonas phi* [Crpm.phi_C; S73904, X56806, X52912], *Ochromonas danica* [Ochr.dan_C; X53183, U09440], *Glycine max* [Glyc.max_C; X06428, X66062], and the cyanelle from *Cyanophora paradoxa* [Cynp.par_C; X52497]), and nonphotosynthetic lineages (*Thermotoga maritima* [Tt_maritim; M21774, M27479], *Bacteroides fragilis* [Bac.fragil; M61006, unpublished], *Chlamydia trachomatis* [Clm.tracho, M59178; *tuf*B:M74221], *Borrelia burgdorferi* [Bor.burg13; L39080, L23125], *Flexistipes sinusarabici* [Fls.sinusa; M59231, X59461], *Bacillus subtilis* [B.subtilis; K00637, M10606, X00007, unpublished], *Shewanella putrefaciens* [She.putre2; X81623, unpublished], and *Thermus thermophilus* [T.themoph27; L09659, X05977]). Bootstrap values indicating the level of support for particular edges were determined under split decomposition, parsimony, and neighbor joining. Phylogenetic estimations were made on uncorrected patterns in the data using SPLITSTREE2.4 (Huson 1998) and PAUP* 4.0d63 (Swofford 1998).

In the following analyses, ‘‘invariant’’ sites are those which are unchanged at a given position in a set of aligned sequences. The subset of invariant sites which are unable to change is termed ‘‘invariable.’’ For the contingency test, sequences are partitioned into three groups—plastids, chloroxybacteria, and nonphotosynthetic taxa. Sites in the alignment are designated type 0 for group r (=1, 2, 3) if all the sequences in group r take the same state at that site. Otherwise, we say the site is of type 1 for group r. Furthermore, we say a site is of type (i, j, k) if it is of type i (=0, 1) for group 1, j (=0, 1) for group 2, and k (=0, 1) for group 3. Let $X(i, j, k)$ denote the $2 \times 2 \times 2$ contingency table, in which i, j, and k each take the values 0 or 1, and in which $X(i, j, k)$ is the number of sites of type (i, j, k). Thus, $X(0, 0, 0)$ is the number of sites which are invariant in group 1, invariant in group 2, and invariant in group 3 (the common state for group 1 may, however, differ from the common state for group 2, etc). The sum of the eight $X(i, j, k)$ values clearly equals the total number of sites. Let $Y(i, j, k) = X(i, j, k)$ for all entries except $(i, j, k) = (0, 0, 0)$, in which case we will set $Y(0, 0, 0) = X(0, 0, 0) - x$, where x denotes the unknown number of invariable sites (sites that are unable to change). Under the null hypothesis that sequences evolve under models of type 1 or 2.1, the variable sites behave independently for variability in the three groups (exactly under the Jukes-Cantor [1969] and Kimura [1981] 2ST and 3ST processes and approximately under other similar models). Thus, we can apply a chi-square test to the Y table with 4 degrees of freedom (since for an $I \times J \times K$ contingency table, the number of degrees of freedom for complete independence is $IJK - I - J - K + 2$) (Christensen 1990). To make the test conservative, we select the value of x which minimizes the chi-square statistic. Then, if the chi-square value is still significant, it will also be significant for any value of x in the model.

For pairwise comparisons of groups of sequences, the following terminology is adopted. Sites that are invariant across both groups are designated type 1. Sites which are invariant in the first group and different but invariant across the second group (e.g., AAAA/GGGG) are designated type 2. Sites which are invariant across the first group but not across the second (i.e., 0, 1) are designated type 3, while those invariant across the second group but not the first (i.e., 1, 0) are designated type 4. Sites which vary in both groups (1, 1) are designated type 5.

The number of variable sites in DNA sequences was estimated using a maximum-likelihood procedure (e.g., Sidow, Nguyen, and Speed 1992; Adachi and Hasegawa 1995; Lockhart et al. 1996). That is, an increasing proportion of sites constant across all taxa was removed from the data until the maximum-likelihood estimate was obtained using the HKY (Hasegawa, Kishino, and Yano 1985) model under PAUP* 4.0d63 (PAUP options: likelihood criteria, empirical base frequencies accepted, transition–transversion ratio estimated, user tree specified, describe trees). Local taxa rearrangements in the user (neighbor joining and optimal parsimony) trees had little effect on the invariable site estimates. We expect our point estimate for invariable sites to be conservative, since, in the presence of covariotide and covarion structure, the estimation procedure will underestimate the proportion of sites that are invariable in some sequences. To illustrate this, consider split S1 for 16S rDNA (fig. 1). The observed proportion of invariant sites across all taxa is 0.5914. The maximum-likelihood point estimate for invariable sites is 0.5431. Under RAS models, similar estimates for the proportion of invariable sites are expected across all taxa and for within-group comparisons. This is not an expectation under covarion/covariotide models, where within-group values should be higher than between-group values. For the plastid group, the estimated proportion of invariable sites is 0.5861. For the chloroxybacteria group, the estimated proportion of invariable sites is 0.6714.

The inequality test of covariotide/covarion structure for two groups of taxa examines the difference be-



chloroxybacteria- plastid split (S1)

	% bootstrap support when all sites included			% bootstrap support after removal of type 3 sites across S1			% bootstrap support after removal of type 4 sites across S1			% bootstrap support after removal of type 3+4 sites across S1			% bootstrap support when only type 1+ jacknifed type 3+4 sites present across S1		
	SD	**P**	**NJ**	**SD**	**P**	**NJ**	**SD**	**P**	**NJ**	**SD**	**P**	**NJ**	**SD**	**P**	**NJ**
16SrRNA	96	98	100	31	88	91	73	83	98	0	47	27	90	69	100
*tuf*A	5	15	19	13	10	7	1	10	6	0	6	0	45	88	99
*tuf*A (no *Glycine*)	21	28	41	15	23	35	6	24	23	6	17	17	88	91	100
*tuf*A (no cyanelle, no *Glycine*)	63	36	74	49	38	54	47	33	63	34	35	41	97	97	100
*tuf*A (no cyanelle)	38	33	51	17	26	37	20	24	33	7	21	9	80	93	100

photosynthetic- nonphotosynthetic split (S2)

	% bootstrap support when all sites included			% bootstrap support after removal of type 3 sites across S2			% bootstrap support after removal of type 4 sites across S2			% bootstrap support after removal of type 3+4 sites across S2			% bootstrap support after removal of type 3+4 sites across S2 or S1		
	SD	P	NJ	SD	P	NJ	SD	P	NJ	SD	P	NJ	SD	P	NJ
16S rDNA	78	99	100	1.5	86	72	69	97	99	0	76	54	0	10	0
*tuf*A	0	44	42	0	39	22	0	34	40	0	36	21	0	12	0

FIG. 1.—The importance of type 3 and type 4 sites on bootstrap support for splits S1 and S2. Bootstrap support was calculated after omitting the sites of the types indicated. Column *E* for S1 shows the bootstrap support using a data set that contained only type 1 sites and a subset of type 3 and type 4 sites sampled randomly without replacement (jacknifed). The total number of type 3 and type 4 sites included was made equal to the number of type 5 sites present in the original data. Comparison of columns *A, D,* and *E* emphasizes the importance of the effect of type 3 and type 4 sites on reconstructed tree structure.

tween the expected number of pattern types and the observed number. Let N_i be the number of variable sites of type i for $i = 1, 2, \ldots, 5$, and $N = N_1 + \ldots + N_5$, the total number of variable sites. Thus, the total number of sites is N plus the number of invariable sites. Now, suppose we have a RAS model, and let p_i $(i = 1, \ldots, 5)$ denote the probability that a variable site is of type i. We claim that

$$p_5 - (p_3 + p_5)(p_4 + p_5) \geq 0. \qquad (1)$$

The proof of inequality (1) is as follows. Let V_1 (respectively, V_2) denote the event that a randomly selected variable site is also variable in $G1$ (respectively, $G2$). Let (random variable) R be the rate at which a randomly selected variable site evolves, let π_t be the proportion of variable sites which are evolving at rate t, and let T be the set of all nonzero rates (i.e., the range of R). Then, by definition,

$$P[V_1 \text{ and } V_2] = \sum\nolimits_{t \in T} \pi_t P[V_1 \text{ and } V_2 | R = t]$$

$$P[V_i] = \sum\nolimits_{t \in T} \pi_t P[V_i | R = t], \qquad i = 1, 2$$

Now, for underlying models such as the Kimura 3ST or 2ST model (Kimura 1981) or the Jukes-Cantor (1969) model, we also have

$$P[V_1 \text{ and } V_2 \mid R = t] = P[V_1 \mid R = t] P[V_2 \mid R = t] \qquad (2)$$

(see Tuffley and Steel 1998), and for more general stationary substitution models, this equality holds approximately. Furthermore, the random variables $P[V_1 \mid R]$ and $P[V_2 \mid R]$ are positively correlated, that is, Cov[P[V_1

Table 1
Pattern Types and X^2 and *P* Values for the Contingency Test

	000	111	100	001	010	110	101	011	Optimal x	χ^2	P Value	Sequence Length
16S rDNA	351	110	44	145	16	9	96	52	320	42.7	<0.001	823
TufA	59	39	6	33	5	1	27	11	55	16.4	0.001	181

$| R], P[V_2 | R]] \geq 0$, since $P[V_1 | R = t]$ and $P[V_2 | R = t]$ are both increasing functions of t. Now, since $\mathrm{Cov}[XY] = E[XY] - E[X]E[Y]$ (where E is the expectation operator), we have, from equation (2),

$$\begin{aligned} P[V_1 \text{ and } V_2] &= E[P[V_1|R]P[V_2|R]] \\ &\geq E[P[V_1|R]] \times E[P[V_2|R]] \\ &= P[V_1]P[V_2], \end{aligned}$$

and, thus,

$$P[V_1 \text{ and } V_2] \geq P[V_1]P[V_2].$$

Furthermore, $p_5 = P[V_1 \text{ and } V_2]$, $p_4 + p_5 = P[V_1]$, and $p_3 + p_5 = P[V_2]$, which thereby establishes inequality (1). Inequality (1) provides a test of a RAS model. Consider the statistic

$$W = N_5 - (N_3 + N_5)(N_4 + N_5)/N.$$

By standard asymptotic techniques (see Serfling 1980), provided N is large, W is approximately normally distributed with mean $N(p_5 - (p_3 + p_5)(p_4 + p_5))$ and variance $N\sigma^2$, where $\sigma^2 = \Sigma_i N_i v_i^2/N - (\Sigma_i N_i v_i/N)^2$ and $v_i = \partial W/\partial N_i$.

By inequality (1), the mean value of W should be nonnegative under a RAS model (type 2.1 or 2.2) and 0 under a type 1 model. Thus, if W takes a large negative score (in comparison to $\sigma\sqrt{N}$), then we can reject such a model in favor of some covariotide/covarion mechanism.

Results and Discussion

Contingency Test

In the contingency test, the null hypothesis is that the sequences follow a type 1 model or a type 2.1 model in which a certain (unknown) subset of sites are unable to vary (and this subset is constant across the tree), while the remaining sites undergo substitution at a constant rate under the usual Markov-style models, such as the Jukes-Cantor and Kimura 2ST and 3ST models. The test was applied to 16S rDNA and TufA sequences, and the results are shown in table 1. The values of χ^2 were 42.7 and 16.4 with 16S rDNA and TufA, respectively. Hence, the null hypothesis was strongly rejected ($P < 0.001$ and $P = 0.001$, respectively). This suggests a type 2.2 RAS or covarion/covariotide model. Relevant to interpretation of this result is figure 1, which suggests that if a type 2.2 model describes evolution of the data, then the patterns contributing most to observed phylogenetic structure (splits S1 and S2) are of a single (or very few) rate class(es). These would describe slowly evolving positions which show no character state changes within either the anciently diverged plastid group or the chloroxybacteria group (type 3 and 4 sites). However, as will be seen with the inequality test, there is no expectation under RAS models for the relatively large number of type 3 and type 4 sites observed in 16S and *tuf*A data.

Inequality Test

Since the test shown in table 1 cannot reject a type 2.2 Markov model in which the sequences have variable sites changing at different rates in favor of a covariotide/covarion model, the inequality test was applied comparing (1) plastids with chloroxybacteria and (2) photosynthetic organisms with nonphotosynthetic ones using 16S rDNA and *tuf*A sequences. In essence, this test compares the observed number of sites which are variable in two groups (N_5) with the expected number $(N_3 + N_5)(N_4 + N_5)/N$, which is calculated from the product of the probability of a given site being variable in group 1 and that of a given site being variable in group 2. Under certain covariotide/covarion models, N_5 will be less than expected, since a site which is varied in group 1 will be less likely to be varied in group 2, whereas under a type 2.2 model, a site which is varied in group 1 is more likely to be varied in group 2. The values of N_1 to N_5 for splits S1 and S2 are shown in table 2. With the 16S rDNA sequences, covariotide structure was clearly demonstrated in both S1 ($W = -12.9$; $\sigma\sqrt{N} = 3.5$) and S2 ($W = -16.8$; $\sigma\sqrt{N} = 2.7$) splits based on point estimates of N_1. With the *tuf*A sequences, covariotide structure was shown clearly for the S2 split ($W = -7.1$; $\sigma\sqrt{N} = 1.97$), but less convincingly for the S1 split ($W = -2.2$; $\sigma\sqrt{N} = 2.3$).

Tree Structure

The dependence of the tree structure on sites of the types shown in table 2 was determined by removing sites of particular types(s) from the data set and then calculating the bootstrap support under split decomposition, parsimony, and neighbor joining for the edges separating (1) plastids from chloroxybacteria (S1) and (2) photosynthetic organisms from nonphotosynthetic ones (S2). The results are shown in figure 1.

Sites for which the character states are the same in one group and varied in the other (types 3 and 4) contribute most of the bootstrap support for the split S1, which separates plastid and chloroxybacterial groups (fig. 1). Very few type 2 patterns occur between these groups (table 2); there are none in the *tuf*A sequences and only three in the 16S rDNA sequences. In trees using 16S rDNA sequences, there are sufficient type 3 and type 4 sites relative to the number of type 5 sites to partition plastids from chloroxybacteria with high bootstrap support. This split occurs despite plastids and chloroxybacteria being site-saturated with respect to

Table 2
Absolute Numbers of Pattern Types in 16S rDNA, *tuf*A (first and second codon positions) and TufA (amino acid) for Splits S1 and S2

	Sequence Length	Type 1[a]	Type 2	Type 3	Type 4	Type 5
S1 (chloroxybacteria–plastid)						
16S rDNA	842	498 (41)	3	150	69	122
*tuf*A (1 + 2 codon positions)	394	229 (21)	0	51	39	75
TufA	197	95	0	32	19	51
*tuf*A, no *Glycine*	394	238	2	40	46	68
TufA, no *Glycine*	197	101	0	26	24	46
*tuf*A, no cyanelle	394	229	0	51	50	64
TufA, no cyanelle	197	95	0	32	23	47
*tuf*A, no cyanelle, no *Glycine*	394	238	2	40	62	52
TufA, no cyanelle, no *Glycine*	197	101	0	26	30	40
S2 (photosynthetic–nonphotosynthetic)						
16S rDNA	823	351 (7)	0	141	70	261
*tuf*A (1 + 2 codon positions)	362	156 (5)	0	66	31	109
TufA	181	59	0	33	12	77

[a] Numbers in parentheses are the estimated numbers of variable sites.

each other at sites that are free to vary (Lockhart et al. 1993). Very weak support is found for the partitioning of plastids from chloroxybacteria with the partial *tuf*A sequences (where the bootstrap values are low before removal of any sites; fig. 1*A*). Examination of figure 1, columns *A*, *D*, and *E*, indicates that it is the relative number of type 5 sites compared to type 3 and 4 sites that reduces bootstrap values for split S1 in the *tuf*A data. The significance of this for phylogenetic inference of plastid origins is discussed later.

In both *tuf*A and 16S rDNA sequences, no type 2 patterns occur to support split S2. Rather, differing distributions of invariant sites are sufficient to partition oxygenic photosynthetic taxa from nonphotosynthetic taxa. (These sites include those which are invariant across the nonphotosynthetic organisms and invariant within either the plastid group or the chloroxybacterial group.) Indeed, the conclusion that most of the support for splits S1 and S2 relies on covariotide patterns of substitution differs from those in earlier analyses (e.g., Delwiche, Kuhsel, and Palmer 1995). However, our results do not suggest that there is necessarily an absence of useful phylogenetic information in these data, since tree structure resulting from covariotide/covarion substitution may be consistent with genuine evolutionary relationships; this will be the case when more closely related groups share a similar distribution of variable sites. For example, the existence of homologous PsbO polypeptides (the extrinsic 33-kDa component of photosystem II) in chloroxybacteria and plastids (Fairweather, Packer, and Howe 1994) supports a single origin of oxygenic photosynthesis. It therefore seems likely that the process of sequence evolution has given rise to phylogenetically informative patterns between oxygenic photosynthetic and nonphotosynthetic groups.

Further data are needed to determine whether the substitution patterns between plastids and chloroxybacteria are phylogenetically informative or misleading and, therefore, to determine whether plastids had a monophyletic or polyphyletic origin. Gene organization data are likely to be of particular value for this (e.g., Douglas 1994). Although at present, such data lend some support to the hypothesis of a monophyletic origin for all plastids, further comparative data are needed from diverse eubacteria to confirm suggested synapomorphic gene arrangements in plastids with different light-harvesting systems.

In a more general context, our results highlight the importance of considering the effect of differing distributions of variable and invariable sites on tree structure, especially when base composition biases are present in sequence data. Our finding of the importance of the effect of covariotide structure on tree shape is significant, as it provides an explanation for why different compositionally biased plastid and chloroxybacterial sequences can join in reconstructed trees (as noted by Delwiche, Kuhsel, and Palmer 1995). Further, in *tuf*A eubacterial sequences, at least for the sequence lengths determined, fewer type 3 and type 4 sites than in 16S rDNA occur between plastid and chloroxybacterial sequences (fig. 1). As a consequence, compositional biases in site-saturated plastid and chloroxybacterial *tuf*A sequences significantly distort local tree shape. Hence, the choice of which compositionally biased *tuf*A sequences are used to build a tree will strongly bias support for or against competing hypotheses of plastid origins. This phenomenon was previously reported for eubacterial *tuf*A sequences (Lockhart et al. 1992) and also for *sec*A (Barbrook 1996). As predicted by the study of Naylor and Brown (1997), biased substitutions in *tuf*A amino acid sequences are most evident from the relative frequencies of aliphatic amino acid residues I and V (but not L) at variable positions in the sequences (e.g., mean frequency ± SD for I and V residues in an alignment of 10 photosynthetic taxa—155 amino acids at varied positions for *Astasia* I [0.194 ± 0.032] and V [0.065 ± 0.0199] and *Glycine* I [0.065 ± 0.0199] and V [0.161 ± 0.030]).

Reliable phylogeny reconstruction will require an understanding of the evolution of molecules such as rRNA and *tuf*A in terms of changing constraints at different sequence positions. Despite the suggestion that covariotide/covarion patterns of substitution in some sequences may be misleading, there is evidence to suggest that even under conditions of site saturation, covariotide/covarion patterns of change in some groups may allow the retrieval of evolutionary relationships at deep phylogenetic levels (Fitch and Markowitz 1970; Miyamoto and Fitch 1995; Philippe et al. 1996; Waddell, Penny, and Moore 1997). Only with more detailed characterization of the substitution processes in sequence data will we be able to distinguish misleading substitution patterns from those which are phylogenetically informative.

Acknowledgments

We thank David Penny, Vince Moulton, Uwe Maier, Sue Douglas, and the anonymous referees for helpful comments and the Alexander von Humboldt Foundation, the New Zealand Marsden Fund, the Wellcome Trust, and the German Ministry of Science and Technology for their financial support of this work. C.J.H. thanks the Institute of Molecular Biosciences, Massey University, for their hospitality during part of this work. We especially thank David Swofford for a prerelease version of PAUP* and Chuck Delwiche and Jeff Palmer for sending us their alignment of *tuf*A.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. **40**:622–628.

BARBROOK, A. C. 1996. Cyanobacterial protein processing and translocation. PhD thesis, University of Cambridge.

CHRISTENSEN, R. 1990. Log-linear models. Springer-Verlag, New York.

DELWICHE, C. F., M. KUHSEL, and J. D. PALMER. 1995. Phylogenetic analysis of *tuf*A sequences indicates a cyanobacterial origin of all plastids. Mol. Phylogenet. Evol. **4**:110–128.

DOUGLAS, S. E. 1994. Chloroplast origins and evolution. Pp. 91–118 *in* D. A. BRYANT, ed. The molecular biology of cyanobacteria. Kluwer Academic Press, the Netherlands.

FAIRWEATHER, M. S., J. C. L. PACKER, and C. J. HOWE. 1994. The extrinsic proteins of photosystem II in photosynthetic organisms: distribution, properties and evolutionary implications. Biochem. Biophys. Res. Comm. **205**:1497–1502.

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

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

HASEGAWA, M., and T. HASHIMOTO. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **21**:160–174.

HUSON, D. H. 1998. SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.

JERMIIN, L. S., P. G. FOSTER, D. GRAUR, R. M. LOWE, and R. H. CROZIER. 1996. Unbiased estimation of symmetrical directional mutation pressure from protein-coding DNA. J. Mol. Evol. **42**:476–480.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MANRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

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

LOCKHART, P. J., C. J. HOWE, D. A. BRYANT, T. J. BEANLAND, and A. W. D. LARKUM. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. **34**:153–162.

LOCKHART, P. J., A. W. D. LARKUM, M. A. STEEL, P. J. WADDELL, and D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA **93**:1930–1934.

LOCKHART, P. J., D. PENNY, M. D. HENDY, and A. W. D. LARKUM. 1993. Is *Prochlorothrix hollandica* the best choice as a prokaryotic model for higher plant chl-*a*/*b* photosynthesis? Photosynth. Res. **37**:61–68.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

MIYAMOTO, M. M., and W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. **12**:503–513.

NAYLOR, G. J. P., and W. M. BROWN. 1997. Structural biology and phylogenetic estimation. Nature **388**:527–528.

OLSEN, G. J., C. R. WOESE, and R. OVERBEEK. 1994. The winds of (evolutionary) change. J. Bacteriol. **176**:1–6.

PALMER, J. D. 1993. A genetic rainbow of plastids. Nature **364**: 762–763.

PESOLE, G., G. DELLISANTI, G. PREPARATA, and C. SACCONE. 1995. The importance of base composition in the correct assessment of genetic distance J. Mol. Evol. **41**:1124–1127.

PHILIPPE, H., G. LECOINTRE, H. L. V. LE, and H. LE GUYADER. 1996. A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. Mol. Biol. Evol. **13**: 1174–1186.

SERFLING, R. J. 1980. Approximation theorems of mathematics statistics. Wiley, New York.

SHOEMAKER, J. S., and W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. **6**:270–289.

SIDOW, A., T. NGUYEN, and T. P. SPEED. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. **35**:253–260.

SWOFFORD, D. L. 1998. PAUP*. Sinaur, Sunderland, Mass.

TUFFLEY, C., and M. A. STEEL. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math. Biosci. **147**:63–91.

WADDELL, P. J., D. PENNY, and T. MOORE. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol. Phylogenet. Evol. **8**:33–50.

ROSS H. CROZIER, reviewing editor

Accepted June 1, 1998