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 tufA 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., Pulmer 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 invariant [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 variable sites, and the variable sites are evolving at different rates.); and type 3—covariotide/covarion models under which sites that are invariant 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...
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 tufA 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 tufA sequences were taken from Delwiche, Kuhlman, and Palmer (1995). RDP loci and GenBank accession numbers, where available for taxa, are given respectively for 16S rDNA and tufA sequences. These were for chloroxy-bacteria (Prochlorothrix hollandica [Prtx.holland; U09445], Gleobacter violaceus [Glb.violac; U09433], Gloethece sp. [Gloh.sp; U09434], Spirulina sp. [Spli.sp2; X75044, X15646], Plectonema boryanum [Plec.borya; U09444], plastids (Astasia longa [Ast.lona.C; X14386, X14385]. Cryphonemash phi [Crpm.phi.C; S73904, X56806, X52912]. Ochromonas danica [Ochr.dan.C; X53183, U09440], Glycine max [Glyc.max.C; X66062, X66062], and the cyanelle from Cyanophora paradoxa [Cynp.par.C; X52497]), and nonphotosynthetic lineages (Thermotoga maritima [Tm.arit; M21774, M27479], Bacteroides fragilis [Bac.fragil; M61006, unpublished], Chlamydia trachomatiss [Clm.tracho, M59178; tuB; M74221], Borrelia burgdorferi [Borburg13; 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 [Tt.thermoph; 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 pairwise distances. Local taxa rearrangements in the tree were 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, described trees). Local taxa rearrangements in the user (neighbor joining and optimal parsimony) trees had little effect on the variable site estimates. We expect our point estimate for variable 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 invariant sites is 0.5431. Under RAS models, similar estimates for the proportion of invariant 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 invariant sites is 0.5861. For the chloroxybacteria group, the estimated proportion of invariant sites is 0.6714.

The inequality test of covariotide/covarion structure for two groups of taxa examines the difference be-
between 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 \( N = N_1 + \cdots + N_5 \), 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

\[
 p_5 = (p_3 + p_5)(p_4 + p_5) > 0. \tag{1}
\]

The proof of inequality (1) is as follows. Let \( V_i \) (respectively, \( V_2 \)) denote the event that a randomly selected variable site is also variable in \( G1 \) (respectively, \( G2 \)). Let \( (\text{random variable}) \) \( R \) be the rate at which a randomly selected variable site evolves, let \( \pi_i \) 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,

\[
 P[V_1 \text{ and } V_2 | R = t] = \sum_{i \in T} \pi_i P[V_i | R = t]
\]

\[
 P[V_i | R = t] = \sum_{i \in T} \pi_i P[V_i | R = t], \quad i = 1, 2
\]

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

\[
 P[V_1 \text{ and } V_2 | R = t] = P[V_1 | R = t]P[V_2 | R = t]
\]

\[
 (2)
\]

(see Tuffley and Steel 1998), and for more general stationary substitution models, this equality holds approximately. Furthermore, the random variables \( P[V_1 | R] \) and \( P[V_2 | R] \) are positively correlated, that is, \( \text{Cov}[P[V_1 | R], P[V_2 | R]] > 0 \).
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 $\chi^2$ 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 tufA 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 tufA 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 $N_5$. With the tufA 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 chloroxybacteria groups (fig. 1). Very few type 2 patterns occur between these groups (table 2); there are none in the tufA 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...
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 tufA sequences (where the bootstrap values are low before removal of any sites; fig. 1A). 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 tufA data. The significance of this for phylogenetic inference of plastid origins is discussed later.

In both tufA 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 invariant 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 tufA 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 tufA sequences significantly distort local tree shape. Hence, the choice of which compositionally biased tufA 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 tufA sequences (Lockhart et al. 1992) and also for secA (Barbrook 1996). As predicted by the study of Naylor and Brown (1997), biased substitutions in tufA 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 tufA 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

### Table 2

<table>
<thead>
<tr>
<th>Sequence Length</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>Type 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 (chloroxybacteria–plastid)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rDNA</td>
<td>842</td>
<td>498</td>
<td>3</td>
<td>150</td>
<td>69</td>
</tr>
<tr>
<td>tufA (1 + 2 codon positions)</td>
<td>394</td>
<td>229</td>
<td>0</td>
<td>51</td>
<td>39</td>
</tr>
<tr>
<td>TufA</td>
<td>197</td>
<td>95</td>
<td>0</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>tufA, no Glycine</td>
<td>394</td>
<td>238</td>
<td>2</td>
<td>40</td>
<td>46</td>
</tr>
<tr>
<td>TufA, no Glycine</td>
<td>197</td>
<td>101</td>
<td>0</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>tufA, no cyanelle</td>
<td>394</td>
<td>229</td>
<td>0</td>
<td>51</td>
<td>50</td>
</tr>
<tr>
<td>TufA, no cyanelle</td>
<td>197</td>
<td>95</td>
<td>0</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td>tufA, no cyanelle, no Glycine</td>
<td>394</td>
<td>238</td>
<td>2</td>
<td>40</td>
<td>62</td>
</tr>
<tr>
<td>TufA, no cyanelle, no Glycine</td>
<td>197</td>
<td>101</td>
<td>0</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>S2 (photosynthetic–nonphotosynthetic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rDNA</td>
<td>823</td>
<td>351</td>
<td>7</td>
<td>141</td>
<td>70</td>
</tr>
<tr>
<td>tufA (1 + 2 codon positions)</td>
<td>362</td>
<td>156</td>
<td>5</td>
<td>66</td>
<td>31</td>
</tr>
<tr>
<td>TufA</td>
<td>181</td>
<td>59</td>
<td>0</td>
<td>33</td>
<td>12</td>
</tr>
</tbody>
</table>

* Numbers in parentheses are the estimated numbers of variable sites.
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 \textit{tuf}A.

LITERATURE CITED


Ross H. Crozier, reviewing editor

Accepted June 1, 1998