Letter to the Editor

How Molecules Evolve in Eubacteria

Peter J. Lockhart, Daniel Huson,† Uwe Maier,‡ Martin J. Fraunholz,‡ Yves Van de Peer,§ Adrian C. Barbrook,‖ Christopher J. Howe,‖* and *Mike A. Steel¶*

*Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand; †Program in Applied and Computational Mathematics, Princeton University; ‡Fachbereich Biologie, Zellbiologie und Angewandte Botanik, Philipps-Universität Marburg, Marburg, Germany; §Fakultät Biologie, Evolutionsbiologie, Universität Konstanz, Konstanz, Germany; ‖Department of Biochemistry and Cambridge Centre for Molecular Recognition, University of Cambridge, Cambridge, England; ¶Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.

A fundamental assumption in building evolutionary trees is that processes of change are constant across the tree of life (Li and Gu 1996; Swofford et al. 1996). Despite this universal view, it is now clear that nucleotide compositions, amino acid compositions (e.g., Lanave et al. 1984; Sueoka 1988; Hasegawa and Hashimoto 1993; Barbrook, Lockhart, and Howe 1998; Forster and Hickey 1999; Lockhart et al. 1999), and, as we demonstrate here for eubacterial sequences, the distribution of sites in sequences that can accept substitutions may change over time.

We investigated anciently diverged eubacterial sequences using a simple linear dissimilarity measure (d_{lcov}) that was sensitive to the type of variable sequence evolution predicted by a covarion/covariotide model (a model of evolution in which the same sequence positions are free to substitute in some taxa but not in others). Since tree-building properties of d_{lcov} differ under covarion/covariotide and rates-across-sites models, d_{lcov} allowed us to test for evidence of covarion/covariotide evolution in eubacterial sequences. Our analyses demonstrated that evolving distributions of variable sites in molecules provide support for deep-branching patterns in phylogenies reconstructed for eubacterial trees of life. This finding joins growing evidence supporting the covarion/covariotide evolution of sequences (Fitch and Markowitz 1970; Lockhart et al. 1996, 1998; Phillippe and Laurent 1998; Germot and Philippe 1999; Lopez, Forterre, and Philippe 1999; Moreira, Guyader, and Philippe 1999; Philippe et al. 2000; Steel, Huson, and Lockhart 2000).

Given two monophyletic groups of taxa, the site patterns found in an alignment of sequences can be described in terms of five classes (Lockhart et al. 1998). Two of these are used in calculating d_{lcov}. Let N_3 denote the number of sites that are unvaried in the first group but varied in the second group, and let N_4 denote the number of sites that are unvaried in the second group but varied in the first. Let N denote the total number of sites. Thus,

$$d_{\text{lcov}} = \frac{N_3 + N_4}{N}$$

is the proportion of sites varied in one group but not the other. We describe exactly the expected value of d_{lcov} under two models—a model in which there is a distribution of rates across sites (RAS), and a covarion-style model of the type described and analyzed recently by Tuffley and Steel (1998). Under this latter model, the following nonlinear dissimilarity measure converges (with increasing sequence length) to an additive measure that is proportional to the evolutionary distance between the groups:

$$d_{\text{cov}} = -\log\left[N_5 - \frac{(N_3 + N_5)(N_4 + N_5)}{N}\right],$$

where N_5 is the number of sites that are varied in both groups.

At variable positions under both the RAS and the covarion-style models, we assume that the underlying mechanism of nucleotide substitution is described by the Kimura 3ST model (or some submodel). The results are expected to be similar under other models of nucleotide substitution but somewhat more difficult to analyze. Under either the RAS or the covarion-style model, the expected value of N_k/N is $p_k - p_{ij}$, where p_k is the probability that the site is varied among the sequences in group $k \in \{i, j\}$, and where p_{ij} is the probability that the site is varied among the sequences in both groups. Thus, if we let e_{ij} denote the expected value of d_{lcov} (under either model), then $e_{ij} = p_i + p_j - 2p_{ij}$. Consequently, under an RAS model, we have:

$$e_{ij} = p_i + p_j - 2\int P[E_i|\lambda]P[E_j|\lambda]\, dF(\lambda) \qquad (1)$$

where $P[E_k|\lambda]$ is the probability that the sequences in group k are varied at a site evolving at rate λ, and the integration is performed with respect to the distribution of rates across sites. Note that if the sites all evolve at the same rate, $p_{ij} = p_i p_j$.

For the covarion-style model described in Tuffley and Steel (1998), lemma 7 of that paper shows that

$$e_{ij} = p_i + p_j - 2p_ip_j - bx_ix_j \exp(-c\tau_{ij}), \qquad (2)$$

where b and c are positive constants (dependent only on the switching rates between ''variable'' and ''invariable'' states under the model), τ_{ij} is the evolutionary distance between groups i and j, and $x_k = P[E_k|\text{var}] - P[E_k|\text{inv}]$, where $P[E_k|\text{var}]$ (respectively, $P[E_k|\text{inv}]$) is

Key words: covarion, covariotide, nonstationarity, split decomposition.

Address for correspondence and reprints: Peter J. Lockhart, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand.

Mol. Biol. Evol. 17(5):835–838. 2000

the probability that a site is varied for the sequences in group $k \in \{i, j\}$, given that it is variable (respectively, invariable) at the root vertex of this group in the underlying tree.

In comparing formulae (1) and (2) for e_{ij} under the two models, we note that equation (1) does not involve the evolutionary distance τ_{ij} between the groups. Hence, under an RAS model, we cannot expect d_{lcov} to extract phylogenetic signal. However, e_{ij} increases monotonically with τ_{ij} for the covarion-style model (eq. 2) and therefore is a (nonlinear) measure of the phylogenetic distance between the groups. Thus, to a first approximation, an expectation is that the d_{lcov} values should fit a star phylogeny under an RAS model. Under a suitable covarion/covariotide-style model (and with τ_{ij} small and monophyletic groups of similar diversity), the expectation is that d_{lcov} will fit the underlying bifurcating tree. We tested if d_{lcov} would allow the recovery of tree shapes similar to the model tree when sequences evolved under a non–covarion/covariotide model. Hence, for sequences of finite length (c = 100, 200, 300, 400, and 500), we simulated the evolution of five groups of sequences (each containing four sequences) on a bifurcating tree under Jukes-Cantor and RAS models (gamma law distribution of rates with shape parameters 0.5, 1, and 1.5), where the numbers of expected substitutions per site were set to 0.2 for the internal edges and to 0.1 for the external ones. For all combinations of parameters, we generated 100 different data sets. To each such data set we then reconstructed splitsgraphs (Bandelt and Dress 1992; Huson 1998) using (1) d_{lcov} and (2) traditional distance measures, corrected according to the model used to simulate the data. Unlike the model transformation, d_{lcov} tended to produce a splitsgraph that did not favor a particular bifurcating tree. Next, we applied d_{lcov} and split decomposition to five different eubacterial tree of life data sets. For the analyses carried out, sequences were sampled from eubacterial groups (e.g., oxygenic photosynthesis, low G+C gram positives, etc.) so as to cover as much of the genetic diversity of each group as possible yet also maintain a hierarchical structure within each group. These steps were carried out in an attempt to identify diverse sequences showing the most conserved group structure. Sequences whose presence produced unresolved trifurcations between basal lineages within groups were excluded, since these perturbed the treelike properties of both d_{lcov} and d_{cov} (i.e., the splitsgraphs became boxlike). Groups that were poorly sampled with shallow divergences were also avoided. The list of taxa used, along with the alignments, are available from http://www.massey.ac.nz/~imbs/Research/MolEvol/Farside/Plants.html.

For each data set, figure 1 shows (1) unweighted bootstrap neighbor-joining trees (obtained using PAUP, version 4; Swofford 1999) recovered using uncorrected (Hamming) distances and (2) split decomposition graphs (obtained using Splitstree, version 3.1; Huson 1997) recovered using d_{lcov}. Since split decomposition makes no assumption that data fit a bifurcating tree, it provides a conservative test for identifying covarion/covariotide

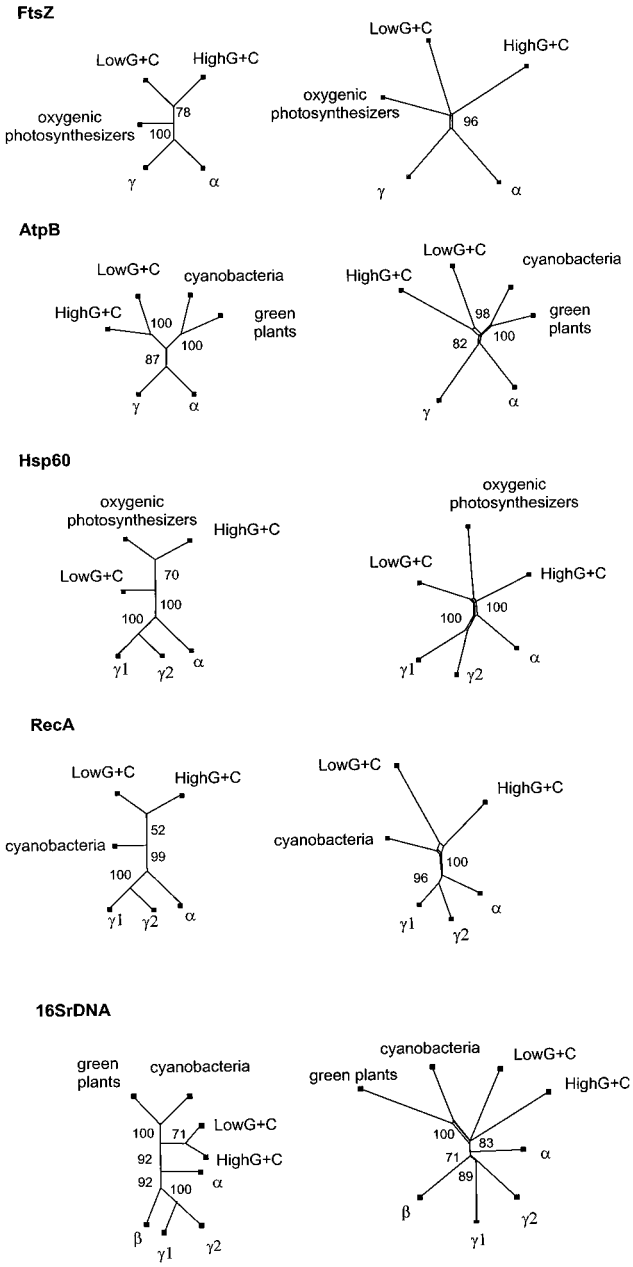

FIG. 1.—Neighbor-joining trees (left) and splitsgraphs (right) for five eubacterial data sets. Total sequence (Hamming) differences were used in construction of the neighbor-joining trees, and group dissimilarity measures were used in reconstructing the splitsgraphs. With protein sequences, $d_{ij} = d_{\text{lcov}} = (N_3 + N_4)/N$. For 16S rDNA, $d_{ij} = (xN_2 + N_3 + N_4)/N$, where weightings for x = 2–4 gave the bifurcating graph shown. Gamma proteobacterial groups 1 and 2 correspond to strongly supported splits in the Hamming distance/neighbor-joining trees.

support for splits which occur in the neighbor-joining trees.

Comparisons of the neighbor-joining trees and splitsgraphs for protein sequences indicate that the distributions of N_3 and N_4 patterns in the different data sets give rise to treelike distances for d_{lcov} and splits that correspond to those recovered most strongly in the neighbor-joining trees (e.g., the splits between the α and γ proteobacteria and between the proteobacteria and oth-

er groups). These observations are explained if sequences belonging to the different monophyletic groups differ in their distributions of variable sites and if these differences provide support for the treelike structures recovered by tree-building algorithms such as neighbor joining.

Less support is provided by $N_3 + N_4$ patterns in the 16S rDNA sequences studied here. With these data, the expected phylogenetic–neighbor- joining 16SrDNA tree is recovered only if we include in our dissimilarity measure an additional pattern class N_2 (i.e., sites at which the character states are different between the two groups and unvaried within each group). The evolution of these patterns is equally well described by covarion and noncovarion models. Thus, with rDNA, while there is evidence for covarionlike patterns of evolution in this molecule (Lockhart et al. 1998), the extent to which these contribute to the inferred phylogenetic relationship between major eubacterial groups is less clear.

It is reassuring that the strongest splits recovered in our protein splitsgraphs reconstructed using d_{lcov} are found with different eubacterial data sets and are also recovered using the nonlinear covarion transform d_{cov} (figures not shown), suggesting a common evolutionary history for these different molecules. However, the extent to which asymmetric processes of change may be convergent (and potentially misleading for phylogeny reconstruction) across more widely sampled groups in trees of life is a question that requires further study. Biased amino acid and nucleotide compositions can be convergent (Barbrook, Lockhart, and Howe 1998; Forster and Hickey 1999; Lockhart et al. 1999), and they are known to cause a problem for phylogeny reconstruction when sequences accepting biased substitutions also share similar distributions of varying sites (Lockhart et al. 1998). Although changes in distributions of variable sites may help to ''fossilize'' phylogenetic history in sequences (Lopez, Forterre, and Philippe 1999), some changes may cause problems for tree building. This can occur if the proportion of variable sites in sequences increases independently in different lineages (e.g., Lockhart et al. 1998; Philippe and Laurent 1998; Germot and Philippe 1999; Steel, Huson, and Lockhart 2000). In this case, the data can be described by the type of inconsistency phenomena discussed by Felsenstein (1978). Such processes have been suggested to mislead outgroup placement with duplicated genes (Lockhart et al. 1996; Philippe and Forterre 1999) and also to mislead the divergence order of eukaryotes (Germot and Philippe 1999; Philippe et al. 2000). These results and those we report here highlight the need for improving our understanding of the biochemical basis for processes of asymmetrical change in sequence evolution. This knowledge would surely help provide confidence in the phylogenetic inference of ancient divergences.

A final point is that we do not propose d_{lcov} as an additive distance measure for building evolutionary trees. The measure is not expected to extract all the useful information present in the sequences, and, as we have pointed out, observations on diverse data sets suggest that the evolution of some sequences occurs by covarion processes which are nonstationary. This is a phenomenon which is difficult to model.

Acknowledgments

We acknowledge support from the Alexander von Humboldt Foundation, the Deutsche Forschungsgemeinschaft, the New Zealand Marsden Fund, the New Zealand/German co-operation agreement, the Broodbank Fund, and the BBSRC.

LITERATURE CITED

BANDELT, H. J., and A. W. M. DRESS. 1992. Split decomposition: a new and useful approach to phylogenetic distance data. Mol. Phylogenet. Evol. **1**:242–252.

BARBROOK, A. C., P. J. LOCKHART, and C. J. HOWE. 1998. Phylogenetic analysis of plastid origins based on SecA sequences. Curr. Genet. **34**:336–341.

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

FITCH, W. F., 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.

FORSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284–290.

GERMOT, A., and H. PHILIPPE. 1999. Critical analysis of eukaryotic phylogeny: a case study based on on the HSP70 family. J. Eukaryot. Microbiol. **46**:116–124.

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

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

LANAVE, C., G. PREPARATA, C. SACCONE, and G. J. SERIO. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20**:86–93.

LI, W. H., and X. GU. 1996. Estimating evolutionary distances between DNA sequences. U.K. edition, London.

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., M. A. STEEL, A. C. BARBROOK, D. H. HUSON, and C. J. HOWE. 1998. A covariotide model describes the evolution of oxygenic photosynthesis. Mol. Biol. Evol. **15**: 1183–1188.

LOCKHART, P. J., C. J. HOWE, A. C. BARBROOK, A. W. D. LARKUM, and D. PENNY. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol. Biol. Evol. **16**: 573–576.

LOPEZ, P., P. FORTERRE, and H. PHILIPPE. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. **49**:496–508.

MOREIRA, D., H. L. GUYADER, and H. PHILIPPE. 1999. Unusually high evolutionary rate of the elongation factor 1a genes from the ciliophora and its impact on the phylogeny of eukaryotes. Mol. Biol. Evol. **16**:234–245.

PHILIPPE, H., and P. FORTERRE. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. **49**:509–523

PHILIPPE, H., and J. LAURENT. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. **8**:616–623.

PHILIPPE, H., P. LOPEZ, H. BRINKMAN, K. BUDIN, A. GERMOT, J. LAURENT, D. MOREIRA, M. MÜLLER, and H. LEGUYADER.

2000. Tree reconstruction and the phylogeny of the eukaryotes. Proc. Natl. Acad. Sci. USA (in press).

STEEL, M. A., D. HUSON, and P. J. LOCKHART. 2000. Syst. Biol. (in press).

SUEOKA, N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85**:2653–2657.

SWOFFORD, D. L. 1999. PAUP. Version 4.65. Sinauer, Sunderland, Mass.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer, Sunderland, Mass.

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

MASAMI HASEGAWA, reviewing editor

Accepted January 18, 2000