Recovering Evolutionary Trees under a More Realistic Model of Sequence Evolution

Peter J. Lockhart, Michael A. Steel,† Michael D. Hendy,† and David Penny**

*School of Biological Sciences and †Mathematics Department, Massey University

We report a new transformation, the LogDet, that is consistent for sequences with differing nucleotide composition and that have arisen under simple but asymmetric stochastic models of evolution. This transformation is required because existing methods tend to group sequences on the basis of their nucleotide composition, irrespective of their evolutionary history. This effect of differing nucleotide frequencies is illustrated by using a tree-selection criterion on a simple distance measure defined solely on the basis of base composition, independent of the actual sequences. The new LogDet transformation uses determinants of the observed divergence matrices and works because multiplication of determinants (real numbers) is commutative, whereas multiplication of matrices is not, except in special symmetric cases. The use of determinants thus allows more general models of evolution with asymmetric rates of nucleotide change. The transformation is illustrated on a theoretical data set (where existing methods select the wrong tree) and with three biological data sets: chloroplasts, birds/mammals (nuclear), and honeybees (mitochondrial). The LogDet transformation reinforces the logical distinction between transformations on the data and tree-selection criteria. The overall conclusions from this study are that irregular A,C,G,T compositions are an important and possible general cause of patterns that can mislead tree-reconstruction methods, even when high bootstrap values are obtained. Consequently, many published studies may need to be reexamined.

Introduction

Conventional tree-building methods from amino acid and nucleotide sequences can be unreliable when the base composition of taxa varies between sequences (Saconne et al. 1989; Penny et al. 1990; Sidow and Wilson 1990; Lockhart et al. 1992*a*, 1992*b*; Forterre et al. 1993; Hasegawa and Hashimoto 1993; Sogin et al. 1993; Steel et al. 1993*b*). The methods tend to group sequences of similar nucleotide composition irrespective of the evolutionary history of the organisms. Ad hoc methods to reduce this problem have been tried but are limited because there has been "no accepted theoretical method for compensating for the effects of biased nucleotide compositions" (Sogin et al. 1993, p. 795). We earlier described a method for measuring, but not overcoming, the problem for small data sets (Lockhart et al. 1993; Steel et al. 1993*b*).

However, we now report a new transformation, LogDet, that allows tree-selection methods to consistently recover the correct tree when sequences evolve under simple asymmetric models that can vary between lineages. Such models produce sequences of different nucleotide compositions (Steel 1993) and in this way are more realistic than most standard models. We show for both theoretical and biological cases (chloroplast origins, bird-mammal relationships, and honeybees) that, where conventional methods select the wrong tree, the LogDet transformation allows the correct phylogeny to be recovered.

Standard evolutionary models are described by stochastic matrices that give the expected rate of change between nucleotides along an edge of the tree. Current tree-building methods implicitly assume a restricted set of matrices, usually time reversible and stationary, to describe the process of change on a tree (for background and examples of the matrices used, see Rodriguez et al. 1990). However, biological data can require different matrices to describe changes in different parts of the tree. With larger distances between taxa, even small deviations from these simple models can mislead existing tree-building methods (Lockhart et al. 1992*a*). The problem with extending standard corrections—e.g., those based on the Jukes-Cantor and Kimura two- and three-parameter models—is that they depend on the multiplication of the matrices being commutative (order independent). Most pairs of matrices do not have this

Key words: amniotes, nucleotide composition, chloroplast origins, determinants, evolutionary models, evolutionary trees, honeybees.

Address for correspondence and reprints: Peter J. Lockhart, School of Biological Sciences, Massey University, Palmerston North, New Zealand.

Mol. Biol. Evol. 11(4):605–612. 1994.

0737-4038/94/1104-0004$02.00

property, and this has limited the majority of evolutionary models to special types of stochastic matrices (Lanave et al. 1984; Hasegawa et al. 1985) where multiplication is commutative.

Models using this restricted set of transition matrices have an advantage of not only recovering a unique tree, but of also providing estimates of objective ("true") lengths (expected number of substitutions) for each edge of that tree. However, these models cannot allow variation of nucleotide frequencies in different lineages, except under restrictive assumptions (e.g., see Bulmer et al. 1991). It has recently been shown (Steel 1993) that under a much more general model there is a method that, without attempting (except in some special cases) to estimate the objective edge lengths, still allows the tree to be recovered. This approach, using logarithms of determinants, will now be described and illustrated with three biological examples.

Methods

Our new LogDet transformation (Steel 1993) bypasses the difficulty mentioned above by using the determinants of the matrices (and multiplication of these, being real numbers, is commutative). For each pair of taxa x and y, we record a "divergence matrix" F_{xy}. This is an $r \times r$ matrix ($r = 4$ for nucleic acid sequences; and $r = 20$ for amino acid sequences), with entries being non-negative and summing to 1. The ijth entry of F_{xy} is the proportion of sites in which taxa x and y have character states i and j, respectively; an example is shown in table 1. For each pair of taxa x and y a single dissimilarity value, d_{xy}, is calculated using the following transformation (Steel 1993)

$$d_{xy} = -\ln[\det F_{xy}], \qquad (1)$$

(where det is the determinant of the matrix, and ln the natural logarithm—hence the name "LogDet"). This approach has fundamental differences from an apparently similar transformation described by Barry and Hartigan (1987). Their measure is based on a different matrix than our F_{xy} and, consequently, may not converge to a treelike metric, since it will not, in general, be symmetric (the dissimilarity between i and j may differ from the dissimilarity between j and i). Nevertheless, the variance of d_{xy} (σ^2_{xy}) can be estimated by techniques similar to those used by Barry and Hartigan (1987). In this case,

$$\sigma^2_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{r} [(F^{-1}_{xy})_{ji}^2 (F_{xy})_{ij} - 1]/c, \qquad (2)$$

Table 1
F_{xy} Multiplied by c

Euglena gracilis Site	*Olithodiscus luteus* Site			
	a	c	g	t
All sites:[a]				
a	224	5	24	8
c	3	149	1	16
g	24	5	230	4
t	5	19	8	175
Parsimony sites:[b]				
a	21	0	7	5
c	0	7	0	6
g	10	3	7	3
t	5	9	7	31

NOTE.—Data are the no. of times a nucleotide (a, c, g, and t) in *E. gracilis* was matched to each nucleotide in the chromophyte *O. luteus*. Thus, in the full sequence, there are 16 sites where *Euglena* had cytosine and *O. luteus* thymine. If parsimony sites alone are examined there are six sites. For calculation these nos. are replaced by the frequencies, to give F_{xy} as defined in the text. The values in F_{xy} differ from those in the Barry and Hartigan (1987) calculation. For each comparison in their matrix the rows are summed and divided by the number of nucleotides; for example, for parsimony sites, the first entry would be $21 \times 33/121$, and the converse comparison (*O. luteus* to *E. gracilis*) would sum the rows, and the first entry would then be $21 \times 36/121$.

[a] $c = 900$ homologous sites of 16S rRNA sequences. The determinant of F_{xy} has a value of 0.002, and $d_{xy} = 6.216$ with $\sigma^2_{xy} = 0.004$.

[b] $c = 121$ parsimony sites. The determinant of F_{xy} has a value of 6.27×10^{-5}, and $d_{xy} = 9.677$ with $\sigma^2_{xy} = 0.849$.

where c is the sequence length. The value of d_{xy} tends to increase with the size of the off-diagonal entries in the divergence matrix.

The d_{xy} transformation allows the correct tree to be recovered but does not estimate the lengths of edges. However, for special models (stationary, with equal nucleotide frequencies) the edge lengths can be obtained with a modification of d_{xy} by adding either $\ln(\det F_{xx}F_{yy})/2$ or $-r \cdot \ln(r)$ and scaling by $1/r$, e.g., setting

$$d'_{xy} = \{d_{xy} + [\ln(\det F_{xx}F_{yy})]/2\}/r, \qquad (3)$$

where F_{xx} and F_{yy} are matrices whose entries give the frequencies of character states for taxa x and y.

Some restricted models have been covered by other authors, including Rodriguez et al. (1990), Tamura (1992), and Bulmer et al. (1991). However the LogDet allows tree reconstruction under much more general conditions than the assumptions described in those papers, requiring only that the determinant of the underlying transition matrices in the tree are not 0, 1, −1. Under the usual independence assumptions (across sites and across the tree) values of d_{xy} (and d'_{xy}) will converge with increasing sequence length, to a treelike metric (satisfying the "four point condition"; Bandelt and Dress 1992). Thus, any "reasonable" tree-selection procedure (such as neighbor joining [Saitou and Nei 1987], split

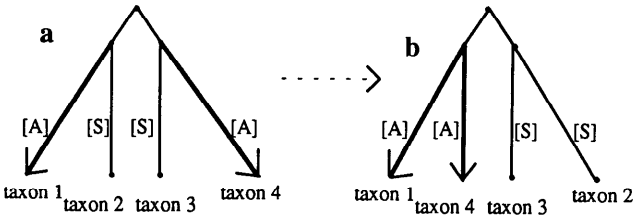

FIG. 1.—Simple stochastic model (i.e., tree, edge lengths, and rates of evolution) that gives sequences with different GC frequencies. The model had both symmetric [S] and asymmetric [A] transition matrices, M_e. An entry $(M_e)_{ij}$ is the probability that the character state at the end of edge e is j, given that it was i at the start of the edge. The matrices were

$$[S]\begin{bmatrix} .91 & .03 & .03 & .03 \\ .03 & .91 & .03 & .03 \\ .03 & .03 & .91 & .03 \\ .03 & .03 & .03 & .91 \end{bmatrix} \text{ and } [A]\begin{bmatrix} .67 & .15 & .15 & .03 \\ .03 & .89 & .05 & .03 \\ .03 & .05 & .89 & .03 \\ .03 & .15 & .15 & .67 \end{bmatrix}.$$

[S] was used on the external edges leading to taxa 2 and 3; and [A] was used on edges leading to taxa 1 and 4. The internal edge used a symmetric matrix with a rate of change 16.7% that of the rate of [S]. Probabilities of all possible sequence patterns were calculated exactly (i.e., they were not simulated) by standard dynamic programing techniques (Smith 1991). Tree-building procedures were tested on these sequences, and all methods using either observed patterns or corrections based on symmetrical transition matrices fail (table 2).

decomposition [Bandelt and Dress 1992], corrected parsimony [Steel et al. 1993*a*], or closest tree [Hendy and Penny 1989]) will converge to the correct tree for sufficiently long sequences generated under this simple model.

Table 2
Results from Analysis of Data Generated under the Model Shown in Figure 1

	0P	1P	2P	3P	LogDet
Parsimony	*b*	*b*	*b*	*b*	*a*
Neighbor joining	*b*	*b*	*b*	*b*	*a*
Split decomposition	*b*	*b*	*b*	*b*	*a*
Closest tree	*b*	*b*	*b*	*b*	*a*

NOTE.—The transformations applied to the data are indicated by the number of parameters used in the correction for multiple changes: 0P = no correction (observed); 1P = Jukes-Cantor; 2P = Kimura two-parameter; and 3P = Kimura three-parameter. Only LogDet corrections led to reconstruction of the correct tree (fig. 1*a*). Maximum likelihood (Felsenstein 1993) was more robust than the other standard methods but still failed. These results emphasize both (1) the important distinction between transformations to the data and the tree-selection criteria (Steel et al. 1993*b*) and also (2) that appropriate mechanisms are required even for selection criteria such as maximum likelihood to be valid. The usual transformations (1P–3P) are based on mechanisms that depend on symmetric divergence matrices that predict that the frequency of each nucleotide will approach equilibrium values of 25% for all taxa, but for biological cases involving widely diverged taxa, when sequences have different proportions of nucleotides, symmetric corrections will seldom accurately describe evolution of the sequences.

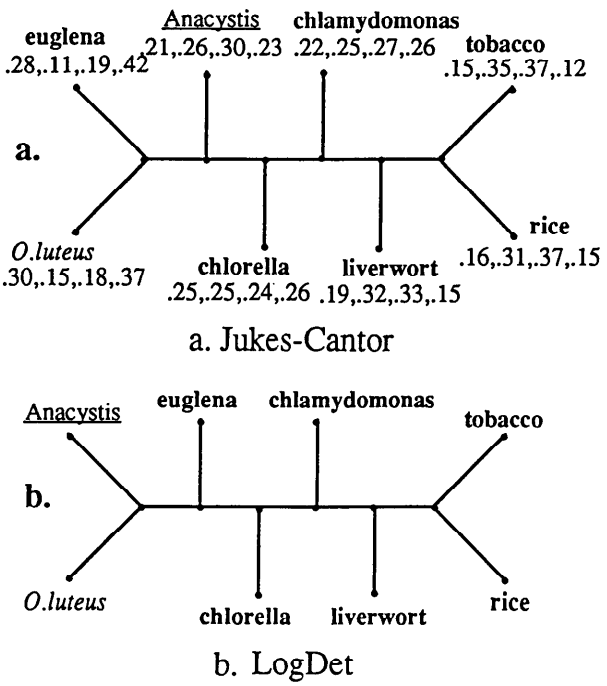

FIG. 2.—Optimal trees found by different procedures for eight photosynthetic taxa. Standard methods produced either tree a (neighbor joining on Jukes-Cantor distances and uncorrected parsimony) or a variant where *Chlorella* and *Chlamydomonas* interchanged (neighbor joining on Kimura two-parameter distances, maximum likelihood [Felsenstein 1993], and a second tree from uncorrected parsimony). The GC contents of the sequences at parsimony sites are shown. When parsimony sites alone were analyzed—and in contrast to the results obtained when Jukes-Cantor corrections were used—the LogDet/neighbor-joining tree places *Euglena* with other chlorophyll *a/b* taxa (*b*). All taxa are photosynthetic, with six having chlorophyll *a/b* light-harvesting complexes, the exceptions being *Olithodiscus luteus* (*italic*), which is a chlorophyll a/c photosynthetic eukaryote, and *Anacystis nidulans* (*underlined*), which has phycobilin accessory pigments. The important difference between trees in a and b is the position of *Euglena*. Sequences are given in Lockhart et al. (1993).

Results

We demonstrate the effectiveness of the LogDet transformation by testing it on theoretical and biological sequences. Figure 1 shows a four-taxon model where two lineages (taxa 1 and 4) have independently acquired a higher GC content. With the stochastic matrices indicated, probabilities of all patterns in sequence data are calculated. These are used to determine which methods recover the original tree. Because the frequencies are for infinitely long sequences, there are no errors introduced by a fixed sample size. Table 2 shows the results from analysis of such data. Only the LogDet correction allows the original tree to be recovered.

Figure 2 shows optimal trees involving a controversial relationship between photosynthetic organelles (Lockhart et al. 1992*a*, 1993) where there are major

Table 3
Euclidean Distances for 18S rRNA Sequences, Based on Nucleotide Frequencies

	Salamander	Frog	Bird	Human	Mouse	Rabbit	Alligator
Salamander		0.1020	0.3359	0.4020	0.4020	0.3826	0.1720
Frog	0.1020		0.3162	0.3499	0.3499	0.3359	0.1720
Bird	0.3359	0.3162		0.1296	0.1296	0.1020	0.1649
Human	0.4020	0.3499	0.1296		0.0000	0.0283	0.2482
Mouse	0.4020	0.3499	0.1296	0.0000		0.0283	0.2482
Rabbit	0.3826	0.3359	0.1020	0.0283	0.0283		0.2245
Alligator	0.1720	0.1720	0.1649	0.2482	0.2482	0.2245	

NOTE.—The frequencies for each nucleotide were calculated for parsimony sites of 18S rRNA (fig. 3), and this information used to calculate the Euclidean distances by using equation (4). No information of sequence order is used in generating this matrix; each sequence used could be randomized and the results would still be the same. Neighbor joining was used with this distance matrix to construct the GC tree (fig. 3*b*). The resulting tree is not interpreted as a "phylogeny" but is simply a test of the extent to which trees built by other methods reflect similarity of nucleotide composition. Salamander is *Ambystoma mexicanum*, bird is *Turdus* species, and frog is *Hyla cinera*.

differences in GC contents between organisms and between nuclear and chloroplast compartments (Lockhart et al. 1992*a*, 1992*b*; Steel et al. 1993*a*). The sequences are for the 16S rRNA of chloroplasts and the cyanobacterium *Anacystis nidulans*. Maximum likelihood, parsimony, and neighbor joining on pairwise distances estimated under the Kimura two-parameter and Jukes-Cantor's one-parameter corrections for all sites in the data were used, and two optimal trees were found. Tree a in figure 2 is one of the optimal trees; the other reversed the positions of *Chlamydomonas* and *Chlorella*. The frequencies of the four nucleotides for each sequence are shown and indicates, for example, that the chromophyte *Olisthodiscus luteus* is most similar in GC content to *Euglena*. However, the placement of *Euglena* closest to the chromophyte breaks up the grouping of chlorophyll *a*/*b* organisms, which all share homologous pigment-binding proteins (Green et al. 1992) and ultrastructural features (Gibbs 1981). This also contradicts trees found from protein sequences (Morden et al. 1992).

When parsimony sites only are analyzed from the 16S data, and the Jukes-Cantor correction is applied, the optimal tree found by neighbor joining also places *Euglena* with *O. luteus*. However, after the LogDet transformation is used at these sites to determine d_{xy} values, the tree selected changes, with the *Euglena* chloroplast sequence now appearing among the other chlorophyll *a*/*b* groups. Tree b in figure 2 shows the optimal tree found by using neighbor joining after LogDet correction, and it links all chlorophyll *a*/*b* taxa. The LogDet transformation has removed the support for an apparently incorrect phylogeny. Although the bootstrap values (not shown) supporting the different hypotheses for the either the Jukes-Cantor- or LogDet-transformed data are not high (most likely because of the age of divergences studied), the results from the LogDet procedure may be preferred both for theoretical reasons (independence of GC content) and because there is now agreement between different classes of data, including other sequence, biochemical, and ultrastructural information. The need to postulate an independent origin of the suite of proteins involved in the chlorophyll *a*/*b* light-harvesting complex is removed.

In the example shown in figure 2 it is easy to identify groupings on the tree that reflect differences in nucleotide composition, but in general it is preferable to have some quantitative measure to detect a grouping of sequences with similar base compositions. One way to do this is to build a tree from a matrix of the Euclidean distances between nucleotide frequencies for each pair of taxa. We call this tree the "GC tree" to indicate that it is based solely on nucleotide frequencies. The tree built using this approach would be the same even if the nucleotides in each sequence were randomly reordered. For each pair of taxa i and j, the Euclidean distance δ_{ij} is given by the formula

$$\delta_{ij}^2 = \sum_k (x_{ik} - x_{jk})^2 , \qquad (4)$$

where x_{ik} is the frequency of nucleotide k = A,C,G, and T for taxon i.

We describe an application for this, with a biological example concerning the relationship between mammals, birds, and crocodilians. Table 3 shows the Euclidean distances, calculated by equation (4) for seven taxa. This matrix was then used by neighbor joining to select the "GC" tree (shown as tree b figure 3). This tree is identical to the tree selected by neighbor joining on both Jukes-Cantor and Kimura two-parameter distances for these 18S rRNA sequences (fig. 3, tree a). This observation is relevant to previous work on the relationship between these species.

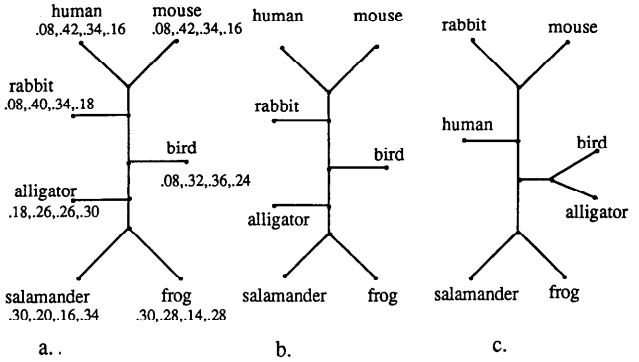

FIG. 3.—Optimal trees for seven vertebrates. The data are aligned 18S rRNA sequences from the rRNA database (Olsen et al. 1991) and GenBank. Tree a is the optimal tree when several standard methods—uncorrected parsimony, maximum likelihood, and neighbor joining—are used with Jukes-Cantor or Kimura two-parameter corrections on all sites. It is identical to the GC tree (tree b), formed by neighbor joining, from the Euclidean distance matrix (table 3), which uses only nucleotide frequency data. Tree c is the neighbor-joining tree after a LogDet correction of the divergence matrix derived from parsimony sites. Bootstrap analyses with neighbor joining (Jukes-Cantor or Kimura two-parameter distances) on either parsimony or all sites supported birds-mammals 99% of the time in tree a and supported birds-crocodilians 96% of the time in tree c.

Studies, particularly with 18S rRNA sequences, have joined mammals and birds as sister groups (Bishop and Friday 1988; Hedges et al. 1990; Rzhetsky and Nei 1992)—rather than birds to crocodilians, as expected on other evidence. Bishop and Friday (1988) pointed out that this result could occur because birds and mammals independently increased in GC content in some chromosome regions (isochores; Bernardi et al. 1985). Although this suggestion has not generally found favor, we have tested it with an 18S rRNA data set obtained from the RDP database (Olsen et al. 1991). Existing methods group the sequences of similar nucleotide composition (fig. 3, trees a and b), but after the LogDet transformation is used, there is strong support, under bootstrap analysis, to join the birds and crocodilians (96% for 500 replicates; fig. 3, tree c). There is clearly an effect of nucleotide composition, since the same tree-selection procedures give different trees, depending on the corrections used for multiple changes.

Our third biological example uses mitochondrial sequences and concerns relationships between six species of *Apis* (honeybee) that are thought to have diverged over the past 40–50 Myr. Parsimony trees were constructed from 500 bootstrap samples taken over all three codon positions. Tree a in figure 4 shows a consensus tree (Felsenstein 1993) for this analysis. This same tree is found when Kimura two-parameter distances are estimated using the same parsimony sites (sometimes, ambiguously, called "informative" sites) and then clustered using neighbor joining. This tree is congruent with both a DNA and an amino acid tree previously found to be optimal under parsimony for these sequences (Willis et al. 1992). However, as pointed out by Willis et al. (1992), this tree contradicts inferences derived from behavioral, morphological, and ecological data. The tree in fact groups taxa of most similar A,G,C,T contents. When information in parsimony sites is transformed using the LogDet method, and the resulting dissimilarity values are clustered with neighbor-joining, the tree obtained (fig. 4, tree c) is congruent with the tree inferred from other biological data.

Discussion

Our conclusion, based on mathematical analysis, simulation, and empirical considerations (congruence between data sets), is that differences in nucleotide composition can mislead current methods but that the LogDet transformation does improve the robustness of tree-selection criteria. Several of our earlier studies demonstrated the potential problems when sequences had unequal nucleotide compositions, and, indeed, with four taxa we could calculate the range of conditions that would lead methods to converge to the wrong tree (Lockhart et al. 1992*b*). We find it important to distinguish between transformations to the data and the tree-selection procedures (Steel et al. 1993*a*); the LogDet procedure is a transformation and not a tree-selection criterion. Sequences have many signals (Penny et al. 1993), including a historical signal, and this new procedure allows the historical signal to be better separated from other signals in the data.

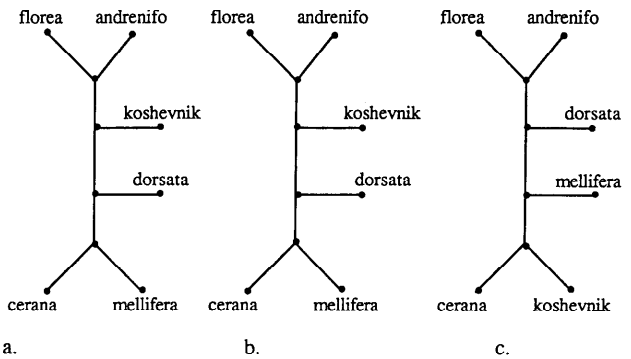

FIG. 4.—Trees for six honeybee species. The data are mtDNA sequences for cytochrome oxidase II (CO II) and are from Willis et al. (1992). Trees are built from 500 bootstrap samples for (tree a) parsimony using all codon positions, (tree b) forming a pairwise distance matrix from the nucleotide frequencies, and (tree c) correcting for multiple changes, with the new LogDet method, on parsimony sites when all codon positions are used. Trees a and b are identical, even though tree b is formed by considering only nucleotide frequencies. The taxa are species of *Apis* with the abbreviated specific names being *A. andreniformis* and *A. koshevnikovi*.

A limitation at present is that, although for simple models the method converges to the correct tree, it generally does not give the amount (or rate) of change along each edge of the tree, except in special restrictive cases. As also recognized with other methods (Shoemaker and Fitch 1989; Sidow et al. 1992), there is still uncertainty as to which sites to use for the correction. Including all sites, particularly for anciently diverged sequences, includes sites that cannot change for functional reasons and consequently results in a serious underestimate of the amount of change. There is also the concern that for any particular site not all compared taxa may be equally free to vary. Table 1 illustrates the two extremes of using all sites and just parsimony sites; the values of d_{xy} are different. This subject requires more exploration.

However, the application of LogDet provides a promising new approach for testing and recovering the tree of life, particularly with regard to the controversies over the deep branches within and between eukaryotes, eubacteria, and "archaebacteria" (Rivera and Lake 1992; Sogin et al. 1993). Inferences from 18S rRNA trees have provided controversy, with the suggestion of two distinct groups within the Eumetazoa (Field et al. 1988). Although it was later suggested that rate inequalities caused erroneous conclusions from these data (Lake 1991), the tree originally derived by these authors reflects the base composition at the parsimony sites of the chosen taxa, and such a tree is not supported under the LogDet transformation. Similarly, the relationship between birds, mammals, and crocodilians was examined recently (Huelsenbeck and Hillis 1993), and it was suggested that unequal rates in different lineages may be the cause of inconsistent inference. However, our results suggest that differing nucleotide frequencies between the compared taxa may be a more serious cause of inconsistency. It is useful to distinguish three usages of the phrase "unequal rates": (1) different rates of evolution (but by the same process) in different lineages, leading to the classic "Felsenstein zone" problem; (2) more generally, different processes in different lineages (leading to the unequal nucleotide frequencies problem discussed here); and (3) variation of rates (or processes) at different sites in the sequence (a further complicating factor).

The results with the honeybee data set are disturbing in that the time of divergence is thought to be within the past 40–50 Myr (Willis et al. 1992). These results illustrate that, even over short periods of divergence (from a geological perspective), A,G,C,T content can affect the amino acid composition of some protein sequences (Crozier and Crozier 1993). Bees have extremely high AT compositions in their mitochondrial genome (Crozier and Crozier 1993), but, nevertheless, to find problems with such recently diverged taxa implies that many published studies should be reconsidered when there are potential effects from differing nucleotide frequencies. This is particularly necessary for anciently diverged taxa, since it follows from our earlier work that even apparently highly conserved sequences may show convergence at the amino acid level (Lockhart et al. 1992*a*, 1992*b*).

As yet we have only three main studies with the LogDet transformation: chloroplasts, birds/mammals, and honeybees. Just because these three studies found effects of unequal nucleotide composition, we cannot generalize to other studies. The three cases were selected because there were contradictions between trees derived from sequences and trees derived from other information. We emphasize that a major use of evolutionary trees is for them to be "predictive" in the sense that a good tree should be an accurate estimator of any results with new data. Too often it appears to be assumed, when there is conflict between data sets, that trees derived from sequences must be correct. It is important to try and resolve the conflicts between data sets, but the results of the present study show that it must not be assumed that the sequences are right and that other information is wrong. Many factors, of which unequal nucleotide frequencies is just one, may need to be considered for a resolution of the conflict.

Another important conclusion is to reemphasize that bootstrap values give no indication as to whether a tree is correct. High bootstrap values indicate that the optimal tree would be unlikely to change as longer sequences become available (convergence), but they give absolutely no indication as to whether the results are converging to the correct tree (consistency) (Penny et al. 1992). The lack of distinction between convergence and consistency is the cause of considerable confusion in many studies. Although it is not a major point of the present study, we find cases (e.g., see fig. 2) where high bootstrap support can be found for different trees, depending on which transformation was used.

Although the LogDet transformation provides researchers with a powerful approach to reconsider existing problems, it is still necessary to look for additional extensions to the LogDet transformation. We are working on extensions that allow variable rates of change at different sites, different weightings for transitions and transversions, and an unbiased estimator that may be more efficient for shorter sequences. Other studies are required to estimate the rate of convergence to a single tree as longer sequences are used. In this study we have illustrated the LogDet transformation with as many as eight taxa, but this is not a limitation. In principle it can be used for the maximum number of taxa that a tree-selection program can use. Recently Lake (1994) and

A. Zharkikh (personal communication) have also independently described measures similar to that given by Steel (1993).

The LogDet transformation is applied here to evolutionary trees, but it is potentially advantageous in other areas of science where asymmetric nonhomogeneous Markov models are used. The LogDet transformation allows biologists to move beyond the simple stationary and/or symmetric Markov models on which Kimura and related correction formulas depend.

Acknowledgments

We thank Ross Crozier, Adrian Gibbs, and two anonymous reviewers for helpful comments on versions of our manuscript. P.J.L. and M.A.S. were supported by a Massey University Research fellowship. Details of program availability can be obtained by e-mail from FARSIDE@massey.ac.nz

LITERATURE CITED

BANDELT, H.-J., and A. DRESS. 1992. A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**: 47–105.

BARRY, D., and J. A. HARTIGAN. 1987. Asynchronous distances between homologous DNA sequences. Biometrics **43**:261–276.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

BISHOP, M. J., and A. E. FRIDAY. 1988. Estimating the interrelationships of tetrapod groups on the basis of molecular sequence data. Pp. 35–58 *in* M. J. BENTON, ed. The phylogeny and classification of tetrapods. Vol. **1**. Clarendon, Oxford.

BULMER, M., K. H. WOLFE, and P. M. SHARP. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88**:5974–5978.

CROZIER, R. H., and Y. C. CROZIER. 1993. The mitochondrial genome of the honey bee *Apis mellifera:* complete sequence and genome organisation. Genetics **133**:97–117.

FELSENSTEIN, J. 1993. PHYLIP 3.5. Available from joe@genetics.washington.edu.

FIELD, K. G., G. J. OLSEN, D. J. LANE, S. J. GIOVANNONI, M. T. GHISELIN, E. C. RAFF, N. PACE, and R. A. RAFF. 1988. Molecular phylogeny of the animal kingdom. Science **239**:748–753.

FORTERRE, P., N. BENACHENHOU-LAFHA, and B. LABEDAN. 1993. Universal tree of life. Nature **362**:795.

GIBBS, S. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic green algae. Ann. N.Y. Acad. Sci. **361**:193–208.

GREEN, B. R., D. DURNFORD, R. ABERSOLD, and E. PICHERSKY. 1992. Evolution of structure and function in the CHL a/b and CHL a/c antenna protein family. Pp. 195–202 *in* N. Murata, ed. Research in photosynthesis. Vol. **1**. Kluwer Academic, Dordrecht.

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. **22**:160–174.

HEDGES, S. B., K. D. MOBERG, and L. R. MAXSON. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

HENDY, M. D., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

HUELSENBECK, J., and D. M. HILLIS. 1993. Success of the phylogenetic methods in the four-taxon case. Syst. Biol. **42**: 247–264.

LAKE, J. A. 1991. Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. Trends Biosci. **16**: 46–50.

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

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.

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

LOCKHART, P. J., D. PENNY, M. D. HENDY, C. J. HOWE, T. J. BEANLAND, and A. W. D. LARKUM. 1992*b*. Controversy on chloroplast origins. FEBS Lett. **301**:127–131.

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. Photosynthesis Res. **73**:61–68.

MORDEN, C. W., C. F. DELWICHE, M. KUHSEL, and J. D. PALMER. 1992. Gene phylogenies and the endosymbiotic origin of plastids. BioSystems **28**:75–90.

OLSEN, G. J., N. LARSEN, and C. R. WOESE. 1991. The ribosomal RNA Database project. Nucleic Acids Res. **19**: 2017–2018.

PENNY, D., M. D. HENDY, and M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. TREE **7**: 73–79.

PENNY, D., M. D. HENDY, E. A. ZIMMER, and R. K. HAMBY. 1990. Trees from sequences: panacea or Pandora's box. Aust. Syst. Bot. **3**:21–38.

PENNY, D., E. E. WATSON, R. E. HICKSON, and P. J. LOCKHART. 1993. Some recent progress with methods for evolutionary trees. N. Z. J. Bot. **31**:275–288.

RIVERA, M. C., and J. A. LAKE. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science **257**: 74–76.

RODRIGUEZ, F., J. L. OLIVER, A. MARIN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. **142**:485–501.

RZHETSKY, A., and M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**:945–967.

SACCONE, C., G. PESOLE, and G. PREPARATA. 1989. DNA microenvironments and the molecular clock. J. Mol. Evol. **29**:407–411.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing trees. Mol. Biol. Evol. **4**: 406–425.

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. NGYEN, and T. P. SPEED. 1992. Capture-recapture. J. Mol. Evol. **35**:253–260.

SIDOW, A., and A. C. WILSON. 1990. Compositional statistics: an improvement of evolutionary parsimony and its deep branches in the tree of life. 1990. J. Mol. Evol. **31**:51–68.

SMITH, D. K. 1991. Dynamic programming: a practical introduction. Ellis Horwood, London.

SOGIN, M. L., G. HINKLE, and D. D. LEIPE. 1993. Universal tree of life. Nature **362**:795.

STEEL, M. A. 1993. Recovering a tree from the leaf colourations it generates under a Markov model. Research rep. 103, May 1993, Mathematics Department, University of Christchurch, N.Z.) Appl. Math. Lett. (in press).

STEEL, M. A., M. D. HENDY, and D. PENNY. 1993*a*. Parsimony can be consistent! Syst. Biol. **42**:581–587.

STEEL, M. A., P. J. LOCKHART, and D. PENNY. 1993*b*. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

WILLIS, L. G., M. L. WINSTON, and B. M. HONDA. 1992. Phylogenetic relationships in the honeybee (genus Apis) as determined by the sequence of the cytochrome oxidase II region of mitochondrial DNA. Mol. Phylogenet. Evol. **1**: 169–178.

SIMON EASTEAL, reviewing editor

Received October 18, 1993

Accepted December 23, 1993