J Mol Evol (1999) 48:264–273

JOURNAL OF MOLECULAR EVOLUTION

A Novel Approach to Phylogeny Reconstruction from Protein Sequences

Nick V. Grishin

Department of Pharmacology, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75235-9041, USA

Received: 17 February 1998 / Accepted: 20 July 1998

Abstract. The reliable reconstruction of tree topology from a set of homologous sequences is one of the main goals in the study of molecular evolution. If consistent estimators of distances from a multiple sequence alignment are known, the distance method is attractive because the tree reconstruction is consistent. To obtain a distance estimate d, the observed proportion of differences p (p-distance) is usually ''corrected'' for multiple and back substitutions by means of a functional relationship $d = f(p)$. In this paper the conditions under which this correction of p-distances will not alter the selection of the tree topology are specified. When these conditions are not fulfilled the selection of the tree topology may depend on the correction function applied. A novel method which includes estimates of distances not only between sequence pairs, but between triplets, quadruplets, etc., is proposed to strengthen the proper selection of correction function and tree topology. A ''super'' tree that includes all tree topologies as special cases is introduced.

Key words: Substitution rates — Amniote phylogeny — Evolutionary distance — Phylogenetic tree

Introduction

The evolutionary distance D between two sequences is usually defined as the number of residue substitutions per site which occur on the shortest path between the two sequences in the tree. The simplest estimate of the distance is the observed proportion of differences between two aligned sequences (the p-distance). This estimate is not consistent, because it misses multiple and back substitutions (Rzhetsky and Sitnikova 1996). Therefore a variety of ''correction'' methods has been proposed (Zuckerkandl and Pauling 1965; Jukes and Cantor 1969; Uzzell and Corbin 1971; Kimura and Ohta 1972; Holmquist et al. 1983; Saitou and Nei 1987; Tajima and Takezaki 1994; Ota and Nei 1994; Grishin 1995; Tourasse and Gouy 1997; Feng and Doolittle 1997; Grishin 1997). The corrected distance $d = f(p)$ is a consistent estimate under the assumed statistical model of sequence change. However, the statistical rules governing the substitution process in real-world sequences remain unknown. Therefore it is problematic to obtain consistent distance estimators.

The consistency of currently available phylogenetic methods has been discussed (DeBry 1992; Steel et al. 1994; Chang 1996a, b). In this article a novel approach to distance method is proposed. First, it is determined if p-distance correction is necessary. Second, in the case that corrections are needed, new methods are described to determine the appropriate correction formula. The implication of the approach is illustrated by three examples.

Method Description

Suppose we have an alignment of homologous sequences. Assume the existence of a twice differentiable correction function $f(p)$, which depends only on p. This correction function should have the following properties.

1. If p is small, then $f(p) \approx p$, since the number of multiple and back substitutions is small. Therefore, assume that $df/dp\ (0) = 1$.

Correspondence to author at current address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; *e-mail:* grishin@ncbi.nlm.nih.gov

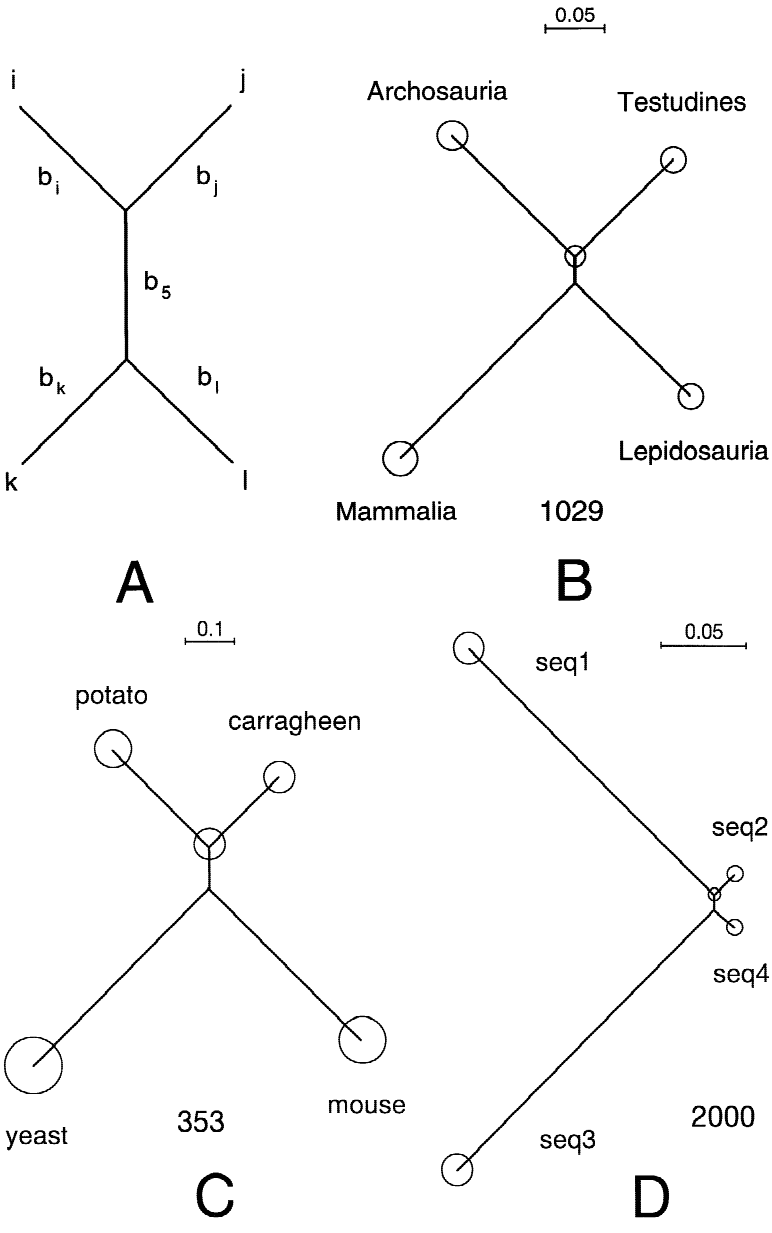

Fig. 1. Phylogenetic trees. **(A)** Unrooted tree of four sequences subject to study in the article, with designations for branch lengths shown. **(B C D)** Phylogenetic trees illustrating text samples 1, 2, and 3, respectively. Tree B shows the amniote relationship on the basis of nine protein families. Tree C is derived from the four cytochrome *b* sequences, and tree D is reconstructed from sequences generated by computer according to the given stochastic model. For each tree the scale bar has the unit of the number of amino acid substitutions per site. Branch lengths are drawn to scale. The radius of a *circle at the tip* of a branch scales with the standard error of the corresponding branch length. The standard error of the length of the middle branch is shown by the *circle at the top* of the branch. The number of sites from which the tree was derived is shown below the tree.

2. $f(p)$ is concave upward, since the rate of accumulation of multiple and back substitutions increases with increasing p. Therefore, assume that $d^2f/dp^2\ (p) > 0$ for $0 \leqslant p < b$, where $b < 1$.

For example, properties 1 and 2 are true[1] for the class of correction formulas in a form $1 - p/b = \int_0^{+\infty} \rho(x) \exp\{-xd/b\}dx$, where d is a distance estimate, $\rho(x)$ is a probability density function of relative substitution rates over sites, and b is the expected value of p for infinitely distant sequences (Ota and Nei 1994; Grishin 1995). It is assumed that for a given alignment b is a constant.

Further consideration is limited to the case of four protein sequences *i, j, k,* and *l.* By $(ij|kl)$ we designate the binary unrooted tree of these four sequences in which sequences *i* and *j* are grouped together. By *i* we designate the branch of the tree with the sequence *i* as a leaf, and b_i is the branch length. The interior branch is designated as 5 with the length b_5 (Fig. 1A). Let d_{ij}^* be an unbiased estimate of the distance D_{ij} between sequence *i* and sequence *j*. For the tree $(ij|kl)$ the four-point condition (Buneman 1974) is true due to the additivity of distances:

$$E(d_{ij}^* + d_{kl}^*) \leqslant E(d_{ik}^* + d_{jl}^*) = E(d_{il}^* + d_{jk}^*) \qquad (1)$$

where $E(x)$ is the expected value of x. Equation (1) can be used to determine the tree topology. In practice one deals with three inequalities in a form

$$d_{ij} + d_{kl} \leqslant d_{ik} + d_{jl} \qquad (2)$$

where d_{ij} is an estimate of the distance D_{ij} between sequence *i* and sequence *j* (see Table 1 for an example). Since unbiased distance estimates are unknown, what is

[1] Subject to some conditions on the function $\rho(x)$.

Table 1. Analysis of *p*-distances[a,b]

Example	Inequality	$S/\sigma(S)$	a
1a	$p_{13} + p_{24} < p_{12} + p_{34}$	0.6	0
	$p_{14} + p_{23} < p_{12} + p_{34}$	**1.3**	**0**
	$p_{14} + p_{23} < p_{13} + p_{24}$	0.7	0
1b	**$p_{13} + p_{24} < p_{12} + p_{34}$**	**2.6**	**0**
	$p_{14} + p_{23} < p_{12} + p_{34}$	1.3	0
	$p_{13} + p_{24} < p_{14} + p_{23}$	1.3	0
2	**$p_{12} + p_{34} < p_{13} + p_{24}$**	**3.0**	**0.14**
	$p_{12} + p_{34} < p_{14} + p_{23}$	2.2	0.17
	$p_{14} + p_{23} < p_{13} + p_{24}$	0.7	0.06
3	$p_{13} + p_{24} < p_{12} + p_{34}$	5.8	1.41
	$p_{12} + p_{34} < p_{14} + p_{23}$	1.4	0
	$p_{13} + p_{24} < p_{14} + p_{23}$	**7.0**	**1.01**

[a] For each example all three inequalities (2) are shown. $S/\sigma(S)$ is the ratio of the difference between the right and the left sides of corresponding inequality to the standard error of this difference, a gives the value of parameter a from the gamma distribution-based correction formula (6), which turns corresponding inequality into equality; a zero value means that inequality never inverts. The line with the largest $S/\sigma(S)$ for each example is in boldface.

[b] In Tables 1, 2, 3, and 4 the sequences are numbered as follows. 1, *Testudines;* 2, *Lepidosauria;* 3, *Archosauria;* 4, *Mammalia* for examples 1a and 1b. 1, carragheen (rhodophyte); 2, potato; 3, yeast; 4, mouse for example 2. 1, seq1; 2, seq2; 3, seq3; 4, seq4 for example 3.

the chance of recovering the correct tree topology when biased distance estimates are used? Will the chances of recovering the tree increase if correction formulas are applied? The partial answer is given in the following theorem.

A Theorem About Invariance to Correction

Let four real values, p_i, $i = 1, 2, 3, 4$, satisfy conditions $0 < p_i < b < 1$. Let the value p_4 be maximal: $p_4 \geqslant p_i$. Let $f(x)$ be a twice differentiable function for $0 \leqslant x < b$. Let $df/dx = 1$ for $x = 0$, and $d^2f/dx^2 > 0$ for $0 < x < b$. Then

$$f(p_1) + f(p_2) < f(p_3) + f(p_4) \qquad \text{if} \qquad p_1 + p_2 < p_3 + p_4$$

For the case $p_1 \leqslant p_i$ or $p_2 \leqslant p_i$ the proof is obvious. If $p_3 \leqslant p_i$, then the proof can be based on Theorem 8 from Bers (1969). The theorem about invariance to correction specifies conditions under which no correction function possessing properties 1 and 2 alters inequality (2). This means that if the maximal among the four observed proportions of differences p appears on the right side of inequality (2), no correction function will invert the inequality. Theoretically if there exists a method to get consistent estimates of distances from the observed proportion of differences via a correction function, then the tree topology is consistently estimated with *p*-distances if the maximal *p*-distance consistently appears on the right sides of all three inequalities (2). One should apply the theorem with care, since due to sampling error the maximal expected *p*-distance may occur on the left side of inequality (2) when the maximal observed proportion of differences is on the right side. The estimate of the standard error of the sum $S = p_{ik} + p_{jl} - p_{ij} - p_{kl}$ shows the statistical significance of inequality (2). The variance of S can be estimated from the variances and covariances of *p*-distances. For a linear function $Y = \sum_{i=1}^{n} a_i y_i$, where a_i are constants and y_i are values of n random variables with the covariance matrix of elements λ_{ij}, the variance of Y is given by the equation $\mathrm{Var}(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \lambda_{ij}$ (Stuart and Ord 1994). A formula, proposed by Bulmer (1991) can be used to estimate the covariance matrix of *p*-distances. Namely, the elements of the covariance matrix for proportion of identical residues are approximated by the equation $\lambda(q_{ij}, q_{kl}) = (q_{ijkl} - q_{ij}q_{kl})/m$, where q_{ijkl} is the proportion of residues identical in all sequences i, j, k, and l, q_{ij} and q_{kl} are the proportions of residues identical in sequences i and j and sequences k and l, respectively, $i < j$, $k < l$, and m is the number of sites without gaps. The case where $i = k$ and $j = l$ gives the variance of q_{ij}, with covariance between q_{ij} and q_{kl} given otherwise. The elements of the covariance matrix for functions of proportions of identical residues $f_i(q_i)$ are approximated by the delta method (Stuart and Ord 1994) as

$$\lambda(f_i(q_i), f_j(q_j)) = \frac{df_i}{dx}(q_i)\,\frac{df_j}{dx}(q_j)\,\lambda(q_i, q_j)$$

The following theorem deals with the case when the maximal value of *p*-distance estimate is on the left side of inequality (2).

A Theorem About Inversion After Correction

Let four real values, p_i, $i = 1, 2, 3, 4$, satisfy conditions $0 < p_i < b < 1$. Let the value p_1 be maximal: $p_1 \geqslant p_i$, and $p_1 \neq p_3$, $p_1 \neq p_4$. Then there exists a function $f(x)$ that is twice differentiable for $0 \leqslant x < b$, $df/dx = 1$ at $x = 0$, $d^2f/dx^2 > 0$ for $0 < x < b$, and

$$f(p_1) + f(p_2) > f(p_3) + f(p_4) \qquad \text{if} \qquad p_1 + p_2 < p_3 + p_4$$

To prove the theorem it is enough to show that for the function $f(x) = ((1 - x)^{-a} - 1)/a$, we have $\lim_{a \to +\infty}\{(af(p_3) + af(p_4) + 2)/(af(p_1) + af(p_2) + 2)\} = 0$. Thus if the maximal among the four observed proportions of differences appears on the left side of inequality (2), the inequality will invert when some correction functions are applied. Since the selected tree topology might be altered if one or more inequalities (2) invert, it is necessary to justify the selection of correction function. Two methods, which facilitate the choice of correction function are proposed.

Inspection of an Inversion Point

For some class of correction functions $f(p, a)$, where a is a parameter, it is possible to find the value a^* that satisfies equation $f(p_1, a^*) + f(p_2, a^*) = f(p_3, a^*) + f(p_4, a^*)$. Inspection of a^* helps to answer whether inequality (2) will invert. For example, consider the correction for-

mula $f(p, a) = ba(\exp\{-\ln\{1 - p/b\}/a\} - 1)$, based on the assumption that the substitution rate varies among site according to the gamma distribution (Uzzell and Corbin 1971; Holmqist et al. 1983). Then for $a < a^*$ inequality (2) inverts. From the analysis of protein sequences it is known that the parameter a usually takes values between 0.5 and 2. Therefore if $a^* = 0.1$, it is unlikely that the inequality inverts. If $a^* = 20$, the inequality is probably inverted. More elaborate schemes of analysis of all three inequalities (2) favoring each of the three topologies can be developed.

Triplet and Quadruplet ''distances''

Triplet, quadruplet, etc., ''distances'' are introduced in this article in addition to the widely used pair distances. The definition of the proportion of identical residues $q_{ij} = 1 - p_{ij}$ in the two sequences i and j extends naturally to the case of 3, 4, . . . , n sequences. Thus the proportion of identical residues in the alignment of n sequences $i_1, \ldots, i_n$ is

$$q_{i_1 \ldots i_n} = \frac{m_{i_1 \ldots i_n}}{m} \tag{3}$$

where $m_{i_1 \ldots i_n}$ is the number of sites that are occupied by the same amino acid type in the sequences $i_1, i_2, \ldots, i_{n-1}, i_n$, $n > 1$, and m is the total number of sites. The definition of the distance between two sequences is extended here to the case of n sequences.

The ''distance'' $D_{i_1 \ldots i_n}$ between n sequences $i_1, \ldots, i_n$ is defined as the number of substitutions per site that occurred on all shortest paths between all pairs of these sequences where each substitution event is counted only once. In other words, if the branch lengths of the tree are proportional to the number of substitutions, the ''distance'' between n sequences is the sum of all branch lengths connecting these n sequences. Thus for the tree of n sequences the ''distance'' between these n sequences is equal to the tree length. For example, in the tree in Fig. 1A $d_{ij} = b_i + b_j$, $d_{ijk} = b_i + b_j + b_k + b_5$, $d_{ijkl} = b_i + b_j + b_k + b_l + b_5$. For the distances defined this way the general formula relating the distance and proportion of identical residues for the pair of sequences extends (see Appendix) for the case of n sequences $i_1, \ldots, i_n$:

$$\frac{q_{i_1 \ldots i_n} - q_\infty^{n-1}}{1 - q_\infty^{n-1}} \approx \int_0^\infty \rho(x) \exp\left\{-\frac{d_{i_1 \ldots i_n} x}{1 - q_\infty^{n-1}}\right\} dx, \qquad n > 1 \tag{4}$$

where $\rho(x)$ is the distribution of relative substitution rates over sites, and q_∞ is the expected proportion of identical residues in a pair of infinitely distant sequences ($q_\infty \geqslant 1/20$ for protein sequences). More exact equations [see Appendix, Eq. A.2] can be used if desired.

Therefore, the ''distance'' between n sequences is estimated readily from the proportion of identical residues in these n sequences, provided that the function $\rho(x)$ is known. Introduction of triplet, quadruplet, etc., ''distances'' enables us to use information which is being lost when only pair distances are considered. Additional observations for statistical estimation of parameters increase the number of degrees of freedom of the system. These ''distances,'' along with the pair distances, can be used to facilitate selection of the appropriate correction function. This leads to the justified selection of the tree topology, and the improvement in estimation of branch lengths of the tree. This can be crucial for the case when selection of the tree topology depends on the correction function. Three methods are proposed for the estimation of a and selection of the tree topology for four sequences.

The Least-Squares Estimation. For each tree topology (μ) let us find the value of parameter a and branch lengths $b_1, \ldots b_5$, which minimize $\| \mathbf{d} - \mathbf{T}_\mu \mathbf{b} \|_2$, where $\mathbf{d}$ is an 11-vector of distances with elements $d_i = f(q_i, a)$, $\mathbf{b}$ is a 5-vector of branch lengths with elements b_j, and $\mathbf{T}_\mu$ is an 11×5-matrix with elements t_{ij}. If the branch length b_j is included in calculations of the distance d_i for topology μ, then $t_{ij} = 1$; otherwise $t_{ij} = 0$. For example, if sequences 1, 2, 3, and 4 are related by the tree with topology (12|34), the vector of distances is $\mathbf{d} = \{d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}, d_{123}, d_{124}, d_{134}, d_{234}, d_{1234}\}$, and the vector of branch lengths is $\mathbf{b} = \{b_1, b_2, b_3, b_4, b_5\}$, then the matrix $\mathbf{T}_{(12|34)}$ is

$$\mathbf{T}_{(12|34)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

The topology for which b_5, the length of the middle branch of the tree, is maximal should be selected. It should be noted that if the middle branch 5 of the tree is longer than some other branches, then $\| \mathbf{d} - \mathbf{T}_\mu \mathbf{b} \|_2$ could be monotonic with a.

Parity Analysis. For unbiased distance estimates d^* the expected value $E(d^*_{1234} - d^*_{12} - d^*_{34})$ is equal to the middle branch length b_5 if the tree has topology (12|34) and is equal to $-b_5$ otherwise. Therefore

$$\begin{aligned} |E(d^*_{1234} - d^*_{12} - d^*_{34})| &= |E(d^*_{1234} - d^*_{13} - d^*_{24})| \\ &= |E(d^*_{1234} - d^*_{14} - d^*_{23})| = b_5 \end{aligned} \tag{5}$$

For a distance estimate $d = f(q, a)$ let us consider the function $g_i(a) = d_{1234} - d_{1i} - d_{jk}$, where i, j, and k are pairwise different integers from the set {2, 3, 4} and $j < k$. Assume that $f(q, a)$ is a decreasing function of a. If the

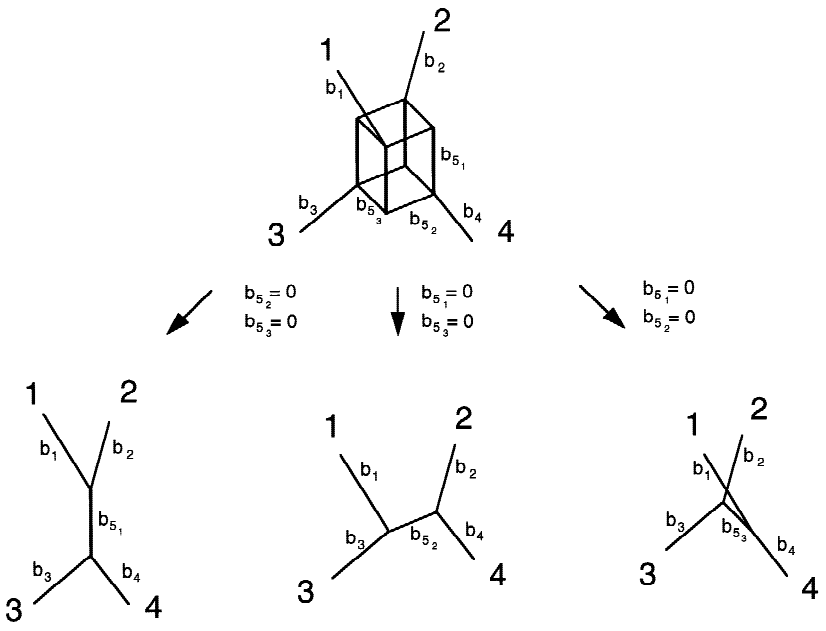

Fig. 2. ''Super'' tree and its special cases. The scheme illustrates how a network of four sequences **(top)** reduces to three trees of different topologies **(bottom).**

function f gives a consistent estimate, then it is likely that g_i is positive for one value of i and negative for the two others. If the distances are severely and nonadditively underestimated (the larger the proportion of differences, the larger the discrepancy between the estimate and the distance), then all three g_i values could be negative. In the case of overestimation more than one g_i value could be positive. If there exist a_i^*, solutions of three equations $g_i(a_i^*) = 0$, then for all a, such that $a_{\min} = \text{mid}(a_2^*, a_3^*, a_4^*) < a < \max(a_2^*, a_3^*, a_4^*) = a_{\max}$, g_i is positive for one value of i and negative for the two others.[2] Let j be the index of the maximal value among three values a_i^*. Then the preferred tree topology is $(1j|ik)$. The optimal value of a for estimation of branch lengths will be the value a^*, which minimizes the function $\sum_{i=2}^{4} (|g_i(a^*)| - \sum_{i=2}^{4} |g_i(a^*)|/3)^2$ on the interval $a_{\min} < a^* < a_{\max}$. The value of $b_5^* = \sum_{i=2}^{4} |g_i(a^*)|/3$ is a topology-independent estimate of the middle branch length, which can be compared to the estimates by the least-squares method for each topology. The statistical hypotheses about their equality can be tested. The favored topology will be the one in which the topology-dependent estimate of the middle branch length matches best the topology-independent estimate.

''Super'' Tree Analysis. Traditionally, the branch lengths are estimated for each tree topology and then the topology, satisfying certain criteria, is selected. However, it is desirable (Yang 1996a) to construct a ''supermodel'' that encompasses all tree topologies as special cases. The network (Fig. 2) is used as a ''super model'' here.[3] In the network of four sequences 1, 2, 3, 4 the vector of branch lengths contains three middle branch lengths (b_{5_1}, b_{5_2}, and b_{5_3}) in addition to $b_1, \ldots, b_4$. If $b_{5_2} = b_{5_3} = 0$, then the network reduces to the tree with topology (12|34). If $b_{5_1} = b_{5_3} = 0$, then the tree has topology (13|24), and if $b_{5_1} = b_{5_2} = 0$, the special case of topology (14|23) arises. Let us find the values of parameter a and branch lengths $b_1, \ldots, b_{5_3}$ which minimize $\| \mathbf{d} - \mathbf{Tb} \|_2$, under the condition $b_{5_1} + b_{5_2} + b_{5_3} - \max(b_{5_1}, b_{5_2}, b_{5_3}) = 0$, where $\mathbf{d}$ is an 11-vector of distances with elements $d_i = f(q_i, a)$, $\mathbf{b}$ is a 7-vector of branch lengths with elements b_j, and $\mathbf{T}$ is an 11×7-matrix with elements t_{ij}. If branch length b_j is included in calculations of the distance d_i, then $t_{ij} = 1$; otherwise $t_{ij} = 0$. For example, if the vector of distances for sequences 1, 2, 3, and 4 is $\mathbf{d} = \{d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}, d_{123}, d_{124}, d_{134}, d_{234}, d_{1234}\}$, and the vector of branch lengths is $\mathbf{b} = \{b_1, b_2, b_3, b_4, b_{5_1}, b_{5_2}, b_{5_3}\}$, then the matrix $\mathbf{T}$ is

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Let $b_{5_i} = \max(b_{5_1}, b_{5_2}, b_{5_3})$; then the topology in which sequences 1 and $i + 1$ are grouped together should be selected.

[2] $\text{mid}(x, y, z) = x$ if $y \leqslant x \leqslant z$ or $z \leqslant x \leqslant y$.

[3] The proposed usage of a network differs from traditional use of network models in evolution studies.

Relations to Other Methods

Traditionally, phylogenetic methods are divided into two groups based on the type of data they use: distance matrix methods and character-state methods (Saitou 1996 and references therein). According to the distance methods, the tree that gives the best fit to the estimated matrix of pairwise distances is chosen. Character-state methods analyze patterns of characters in each site of a multiple alignment. Maximum-parsimony and maximum-likelihood method are character-state methods. For nucleotide sequences the maximum-likelihood method is considered the most efficient, since it uses all the data and statistical models of substitutions on a four-character alphabet are well developed (Schadt et al. 1998; Gu et al. 1995; Tateno et al. 1994; Yang 1993, 1994, 1996; Yang et al. 1994; Felsenstein 1981). However, for protein sequences implementation of maximum likelihood on 20 or even 64 character alphabet, in combination with variability of rates among sites, is limited by computational time, on the one hand, and by sparse data and underdeveloped statistical models of protein sequence evolution, on the other (Felsenstein 1996). Therefore distance methods are widely used for protein sequences (Saitou and Nei 1987). Distance methods do not use all the information contained in a multiple sequence alignment; they consider only pairwise alignments. Therefore, in the currently proposed approach, a compromise between the simplicity of a distance method and the comprehensiveness of a character-state method is made. The *k*-sequence distances are estimated from the *k*-sequence alignments. The proposed approach still misses information, since only patterns of invariant residues are used. In the type of data used the current method is similar to the Hadamard spectral analysis (Hendy et al. 1994) developed for nucleotide sequences. Consideration of all patterns requires a better understanding of substitution processes in protein sequences in combination with variability of rates and patterns of substitutions among sites. Likelihood calculation is crucially dependent on the underlying statistical models. Since the validity of an implemented model is not apparent, it is desirable to use more general assumptions. Therefore, in the proposed approach the conditions under which correction of *p*-distances will not change the tree topology are specified. If these conditions are not fulfilled, the topology selection will depend on the model of variability of rates among sites. The advantages of the presently proposed method over the existing methods can be summarized as follows.

- It allows separation of cases in which the tree topology will not depend on the assumed model of variability of rates among sites.
- It improves the distance method introducing *k*-sequence distances. Consideration of triplet and quadruplet distances in addition to pairwise distances allows for the robust selection of the model of variability of substitution rates among sites. This selection is impossible in a four-sequence case with only pairwise distances.
- It allows estimation of branch lengths of the tree and of a parameter from the distribution of substitution rates over sites, without elaborating statistical assumptions to the level necessary for maximum-likelihood methods that consider all patterns of characters.

Additionally, most tree-building methods (including the maximum-likelihood method) estimate parameters for every tree topology and then select the topology that fits best according to some criteria. Alternatively, in the current method it is proposed to use a network to estimate parameters once and then reduce the network to a tree, eliminating branches whose lengths differ from zero insignificantly.

Examples

The following examples illustrate the proposed methods for the case of four sequences. Designate $v_{1\ldots n} = (q_{1\ldots n} - q_\infty^{n-1})/(1 - q_\infty^{n-1})$, and $b_n = 1 - q_\infty^{n-1}$. It is assumed in calculations that $q_\infty = 1/20$. Three classes of correction functions are analyzed for each example. First,

$$d_{1\ldots n} = b_n a\left(v_{1\ldots n}^{-\frac{1}{a}} - 1\right) \tag{6}$$

which is based on the assumption that the substitution rate varies among sites according to the gamma distribution (Uzzell and Corbin 1971; Holmquist et al. 1983; Ota and Nei 1994; Grishin 1995). The second is

$$d_{1\ldots n} = b_n \frac{1-\beta}{\ln\beta} \ln \frac{\beta - \beta^{1-v_{1\ldots n}}}{\beta - 1} \tag{7}$$

which is suggested from the analysis of spatial structures (Grishin 1997). The two functions given by Eqs. (6) and (7) are among the simplest single-parameter relations that transform the interval [0,1] (the fraction of unchanged residues is defined on this interval) into the interval [0,∞) (evolutionary distance is defined on this interval). The expressions $v^{-1/a} - 1$ and $-\ln\{(\beta - \beta^{1-v})/(\beta - 1)\}$ perform the interval transformation. The expressions a and $(1 - \beta)/\ln\beta$ are scaling factors that allow direct comparison of distances calculated by the different formulas.

The third function is a numerical solution for d of Eq. (4) for the case of the log-normal distribution (Olsen 1987),

$$\rho(x) = (x\sqrt{2\pi c})^{-1} \exp\left\{-\frac{(\ln x + c/2)^2}{2c}\right\} \tag{8}$$

Table 2. Parsimony and least-squares analysis[a]

Example Topology	1a			1b			2			3		
	12|34	13|24	14|23	12|34	13|24	14|23	12|34	13|24	14|23	12|34	13|24	14|23
m.n.s.	109	109	109	621	**616**	620	374	378	**373**	1484	**1480**	1485
a	0.7	0.8	0.7	0.7	0.7	0.6	1.1	1.0	1.1	0.6	0.5	0.7
res	0.11	0.11	0.08	0.03	0.03	0.03	0.07	0.12	0.13	0.02	0.01	0.02
b_5	−0.03	0.01	**0.07**	0.005	**0.025**	−0.018	**0.09**	−0.05	−0.005	**0.009**	−0.022	0.002
$b_5/\sigma(b_5)$	−0.9	0.3	**1.5**	0.5	**2.9**	−2.5	**2.6**	−1.5	−0.2	**1.9**	−4.5	0.8
β	15.	11.	16.	16.	16.	20.	11.	11.	9.	17.	33.	16.
res	0.12	0.12	0.09	0.04	0.03	0.04	0.08	0.15	0.16	0.02	0.01	0.02
b_5	−0.03	0.01	**0.07**	0.006	**0.027**	−0.019	**0.11**	−0.05	−0.001	**0.009**	−0.023	0.002
$b_5/\sigma(b_5)$	−0.7	0.4	**1.5**	0.6	**2.6**	−2.0	**2.4**	−1.2	−0.0	**2.9**	−3.5	0.8
c	1.1	0.9	1.3	1.3	1.4	1.5	1.1	1.0	0.9	1.4	1.8	1.4
res	0.12	0.12	0.10	0.04	0.03	0.04	0.08	0.14	0.15	0.02	0.01	0.02
b_5	−0.02	0.02	**0.07**	0.006	**0.028**	−0.019	**0.11**	−0.04	0.00	**0.009**	−0.023	0.002
$b_5/\sigma(b_5)$	−0.6	0.5	**1.5**	0.7	**2.7**	−1.9	**2.4**	−1.1	0.0	**2.6**	−3.3	0.6

[a] For each example for each topology for each correction function, the best-fit value of the parameter of the correction function [*a* for Eq. (6), β for Eq. (7), and *c* for Eq. (8), least-squares sum of residuals (res), middle branch length (b_5), and ratio of the middle branch length to its standard error [$b_5/\sigma(b_5)$] are shown. m.n.s., minimal number of substitutions (maximum parsimony). The maximal b_5 and $b_5/\sigma(b_5)$ and minimal m.n.s. are in boldface for each example.

A computer program in C language was written by the author to perform the calculations. It runs on a DEC-alpha computer under UNIX. The LAPACK library (Anderson et al. 1995) was used for least-squares calculation, matrix inversion, and SVD. The analysis of all examples follows the same general scheme.

1. Each of the three inequalities (2) for *p*-distances is analyzed. It is determined if the inequality can invert when the correction function is applied, and the range of values of parameter *a* in the gamma distribution-based Eq. (6) for which inequality (2) inverts is found. The statistical significance of the inequality is estimated by the calculation of the standard error of $S = p_{ik} + p_{jl} - p_{ij} - p_{kl}$ (Table 1).
2. For each of the three tree topologies and each of the three correction functions least-squares estimates of the function parameter and tree branch lengths are found (Table 2).
3. For each of the three correction functions the tree topology, suggested by parity analysis, is selected and estimates of the function parameter are found (Table 3).
4. For each of the three correction functions the ''super'' tree is analyzed (Table 4).

Example 1. The Turtle Enigma (Fig. 1B)

The position of turtles (*Testudines*) in the phylogenetic tree of amniotes is highly controversial (Caspers et al. 1996; Rieppel and deBraga 1994; Hedges 1994). Traditionally they are placed to branch before *Lepidosauria* (tuatara, lizards, and snakes) (Eernisee and Kluge 1993). Recently the turtle puzzle became ''hot,'' with several publications in *Nature* suggesting that *Testudines* might have separated from the common ancestor after *Lepidosauria* (Lee 1997; Platz and Conlon 1997; Wilkinson et al. 1997), making turtles advanced diapsid reptiles. For phylogeny reconstruction it is usual to take a large protein family, for example, hemoglobin α. The results of an analysis of homoglobin α sequences from turtle, tuatara, alligator, and human are presented in Tables 1, 2, 3, and 4, example 1a. None of the correction formulas invert inequalities (2), and none of the topologies can be statistically supported (the largest *S* is only about 1.3 of its error). Thus more sites should be added to analysis. Nine protein families containing sequences from all four taxa (*Testudines*, *Lepidosauria*, *Archosauria*, and *Mammalia*) were found in data banks,[4] and sequences were combined. The number of sites increased from 141 in hemoglobin α to 1029 in all nine families. As illustrated in Tables 1, 2, 3, and 4, example 1b, the largest *S* and the middle branch length, which are about three times their error, support grouping *Archosauria* and *Testudines* together. If nine families are analyzed separately (data not shown), six of them favor grouping *Archosauria* with *Testudines* (with the largest middle branch length about 2.5 of its error in myoblobin). The remaining three fami-

[4] For nine protein families the lists of four sequence IDs (Entrez, http://www.ncbi.nlm.nih.gov/Entrez/) for sequences from *Testudines*, *Lepidosauria*, *Archosauria*, and *Mammalia* (human), respectively, follow. Hemoglobin α chain, 1708121, 122487, 122344, 122412; hemoglobin β chain, 1518804, 632037, 2144728, 122615; myoglobin, 70575, 127700, 127633, 21444731; cytochrome *b*, 2147229, 1209488, 117847, 117863; cytochrome *c*, 65465, 118039, 117970, 117996; insulin, 400062, 85933, 124540, 124617; α-crystalline chain A, 1223847, 71478, 71477, 1706112; androgen receptor, 1703693, 1195596, 2134448, 113830; estrogen receptor, 1703692, 1195592, 119597, 2134678.

Table 3. Parity analysis[a]

Example	1a			1b			2			3		
Parameter	*a*	β	*c*	*a*	β	*c*	*a*	β	*c*	*a*	β	*c*
$i = 2$	0.4	108.	3.2	0.6	25.	1.7	**1.3**	**9.**	**0.9**	**0.7**	**15.**	**1.3**
$i = 3$	0.5	54.	2.4	**0.8**	**13.**	**1.2**	0.7	36.	2.0	0.3	359.	3.6
$i = 4$	**0.7**	**19.**	**1.5**	0.4	53.	2.3	0.8	24.	1.6	0.6	21.	1.6
min	0.6	37.	2.1	0.6	21.	1.6	1.0	16.	1.3	0.6	19.	1.5
b_5	0.06	0.07	0.09	0.02	0.02	0.03	0.09	0.11	0.11	0.01	0.01	0.01
$b_5/\sigma(b_5)$	**1.4**	1.2	1.2	**2.4**	2.2	2.2	**2.3**	1.9	1.9	2.9	**3.8**	3.6

[a] For each example for each correction function [*a* for Eq. (6), β for Eq. (7), and *c* for Eq. (8), designate any of these parameters ξ], for each $i = 2, 3, 4$, the solution ξ of the equation $g_i(\xi) = 0$ is given. For each example and for each correction function min gives the value of parameter ξ* that minimizes $\sum_{i=2}^{4}(|g_i(\xi^*)| - \sum_{i=2}^{4}|g_i(\xi^*)|/3)^2$. The middle branch length is estimated as $b_5 = \sum_{i=2}^{4}|g_i(\xi^*)|/3$. The ratio of the middle branch length to its standard error [$b_5/\sigma(b_5)$] is shown. The values of parameters for the favored topology (1*i*|*jk*) and the largest value of $b_5/\sigma(b_5)$ are in boldface.

Table 4. ''Super'' tree analysis[a]

Example	1a			1b			2			3		
Parameter	*a*	β	*c*	*a*	β	*c*	*a*	β	*c*	*a*	β	*c*
param	0.5	70.	2.7	0.5	32.	1.9	0.8	24.	1.6	0.7	14.	1.3
res	0.11	0.15	0.20	0.03	0.03	0.04	0.10	0.14	0.14	0.02	0.02	0.02
b_{5_1}	−0.02	−0.02	−0.03	−0.01	−0.01	−0.01	**0.092**	**0.124**	**0.126**	**0.008**	**0.008**	**0.009**
$b_{5_1}/\sigma(b_{5_1})$	−0.4	−0.2	−0.3	−0.9	−0.7	−0.7	**2.2**	**1.5**	**1.4**	**2.2**	**1.9**	**1.9**
b_{5_2}	0.02	0.02	0.03	**0.024**	**0.027**	**0.030**	−0.02	−0.02	−0.02	−0.00	−0.00	−0.00
$b_{5_2}/\sigma(b_{5_2})$	0.4	0.2	0.3	**2.3**	**1.7**	**1.8**	−0.4	−0.3	−0.3	−0.7	−0.5	−0.5
b_{5_3}	**0.07**	**0.09**	**0.12**	0.008	0.009	0.010	0.015	0.020	0.021	0.004	0.004	0.004
$b_{5_3}/\sigma(b_{5_3})$	**1.6**	**0.9**	**0.9**	1.0	0.7	0.7	0.4	0.3	0.3	1.1	0.9	0.9

[a] For each example for each correction function [*a* for Eq. (6), β for Eq. (7), and *c* for Eq. (8)] the value of the parameter (param), least-squares sum of residuals (res), and values of b_{5_i} and $b_{5_i}/\sigma(b_{5_i})$ for $i = 1, 2, 3$ are given. The maximal b_{5_i} and $b_{5_i}/\sigma(b_{5_i})$ are in boldface.

lies (androgen receptor, cytochrome *b,* and hemoglobin α) favor grouping *Archosauria* with *Lipidsauria* (largest middle branch length 1.6 of its error for the androgen receptor, which is an overestimation, since the number of sites and the number of substitutions are small).

In summary, the data presented here suggest that the turtle problem can be solved with just *p*-distances, provided that the number of sites used in the analysis is large enough. About 1000 amino acid sites combined from nine protein families under the proposed analysis scheme contradict the classical view on the turtle origin and suggest (Fig. 1B) that turtles branched off after *Lepidosauria* and are diapsids.

Example 2. The Minimal Tree Fails (Fig. 1C)

The example with four cytochrome *b* sequences[5] shows that the maximum-parsimony tree (minimum number of substitutions) fails to group sequences from green plant and rhodophyte together (Table 2, example 2). Analysis of the sequences by the methods discussed in this article statistically support the traditional and most probable view (Leblanc et al. 1995; Kumar and Rzhetsky 1996) (Tables 1, 2, 3, and 4 example 2; Fig. 1C). In this example all inequalities (2) invert for very small values of parameter *a*. These values of *a* correspond to an unrealistically high variability of substitution rates over sites. Estimates of the parameter *a* according to the least-squares, parity, and ''super'' tree analysis from the data are much larger (Tables 2, 3, and 4, example 2). Thus in this example *p*-distances are successful again in selection of the tree topology.

Example 3. When Correction Is Crucial (Fig. 1D)

The sequences (seq1 to seq4) used in this example were randomly generated according to a tree with the branch lengths $b_1 = 0.2$, $b_2 = 0.02$, $b_3 = 0.2$, $b_4 = 0.02$, and $b_5 = 0.01$ and topology (12|34). The sequence length was 2000 amino acids. All amino acids were assumed to be equally changeable, but the substitution rates over sites varied according to the exponential distribution $\rho(x) = \exp(-x)$. No gaps were allowed. The generated sequences were analyzed by the methods proposed in this article to test whether the known phylogeny between them is recovered. Analysis of *p*-distances (Table 1, example 3) statistically supports a false group-

[5] Sequence IDs (Entrez, http://www.ncbi.nlm.nih.gov/Entrez/) for the sequences analyzed are 1345906, 231953, 117899, and 117870.

ing of seq2 with seq4, because these sequences are most similar to each other. The same conclusion comes from the parsimony analysis (Table 2, example 3). Thus *p*-distances and maximum parsimony fail to recover the correct topology. Application of the theorem about inversion after correction shows that some inequalities (2) invert when some correction functions are applied (Table 1, example 3). The distances obtained by the best-fit correction function (Tables 2, 3, and 4, example 3) allow us to recover and statistically support the correct phylogeny (Fig. 1D). This example illustrates the suitability of the proposed methods in the case where substitution rates are drastically unequal between lineages, as well as sites.

In summary, the methods of analysis proposed in this article recover reasonable trees in all three examples. In the last two examples they outperform the popular parsimony analysis, which appears misleading. In the examples above three one-parameter correction functions were considered. Their difference is due to the different underlying distributions of substitution rates over sites. The striking feature is that the resulting selection of the tree topology and branch lengths are relatively independent of the type of correction function applied. The three examples illustrate a general tendency: it is not very important which single-parameter distribution of substitution rates over sites is chosen. It is crucial, however, that variations of the parameter in the distribution allow for transition from the case where the substitution rate is highly variable among sites to the case of equal rates for all sites.

Appendix. Relations Between Proportion of Identical Residues and Evolutionary Distance

We define a site to be unchanged over a tree branch i connecting sequences s_1 and s_2, if no amino acid substitutions occurred at this site between sequence s_1 and sequence s_2. The definition implies that the unchanged site is occupied by the same amino acid type in sequences s_1 and s_2. The opposite is not true due to the possibility of ''back'' and ''convergent'' substitutions. We define a site to be unchanged in n sequences, if it was unchanged over all branches of the tree connecting these n sequences.

We assume that sites mutate independently. The distribution of substitution rates over sites remains constant. The probability that a site with a relative substitution rate x remains unchanged over branch i of length b_i is $\exp\{-xb_i\}$. Since the substitutions in different branches occur independently, the probability that the site remains unchanged over branches $1, \ldots, n$ is $\exp\{-x\sum_{i=1}^{n} b_i\}$. Thus the expected number of unchanged sites $u_{1\ldots n}$ in n sequences and the sum of all branch lengths connecting these sequences are related via equation

$$u_{1\ldots n} = \int_0^\infty \rho(x) \exp\{-x \sum_{i=1}^{2n-3} b_i\}\, dx = \int_0^\infty \rho(x) e^{-xd_{i\ldots n}} dx \tag{A.1}$$

where $\rho(x)$ is a probability density function of relative substitution rates over sites. We see that under the present model the functional relationship $u_{1\ldots n} = F(d_{1\ldots n})$ exists. Assume that the conditional probability of the ''back'' substitution provided that a substitution occurred in r. For protein sequences $r \approx 1/19$. Let $u_{1\ldots n}$ be the fraction of unchanged sites in a group of n sequences separated by the ''distance'' $d(1 + r)$. Let $q_{1\ldots n}$ be the fraction of identical sites in a group of n sequences separated by the ''distance'' d. The following equations hold for sequences i, j, k, and l ($n \leqslant 4$):

$$\begin{aligned}
(1+r)q_{ij} &= u_{ij} + r \\
(1+r)^2 q_{ijk} &= (1-r)u_{ijk} + r(u_{ij} + u_{ik} + u_{jk}) + r^2 \\
(1+r)^3 q_{ijkl} &= (1-r)^2 u_{ijkl} + r(1-r)(u_{ijk} + u_{ijl} + u_{ikl} + u_{jkl}) + r^2(u_{ij} \\
&\quad + u_{ik} + u_{il} + u_{jk} + u_{jl} + u_{kl}) + rF((d_{ijkl} - b_5)(1+r)) + r^3
\end{aligned} \tag{A.2}$$

where d_{ijkl} is the ''distance'' between sequences i, j, k, and l and b_5 is the branch length of the middle branch 5 of an unrooted tree, relating four sequences (Fig. 1A). Expressions (A.2) can be derived as solutions of differential equations describing the changes of expected values of proportions of identical residues (N.V. Grishin, unpublished). Equation (4) in the text is an approximation of Eqs. (A.2).

Acknowledgments. The author is grateful to Dr. Keith Henderson for the help with the LAPACK package, numerous computer-related questions, and critical reading of the manuscript, to Dr. Vyacheslav N. Grishin for the help with mathematical issues, to Dr. Vladislav S. Markin and Dr. Arcady R. Mushegian for various discussions, to Dr. Emile Zuckerkandl for helpful suggestions, discussions, and encouragement, to Dr. Hong Zhang and Steve Lockless for critical reading of the manuscript, and to Dr. Ziheng Yang and an anonymous reviewer for helpful comments. This work was partially supported by a grant from the Welch Foundation (I-1257) to Dr. Margaret A. Phillips.

References

Anderson E, Bai Z, Bischof C, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Ostrouchov S, Sorensen D (1995) LAPACK, user's guide, 2nd edition. Society for Industrial and Applied Mathematics, Philadelphia

Bers L (1969) Calculus. Holt, Rinehart and Winston, NY, Chicago, San Francisco, Atlanta, Dallas, Montreal, Toronto, London, Sydney, pp 203, 224–225

Bulmer M (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol Biol Evol 8: 868–883

Buneman P (1974) A note on the metric properties of trees. J Combin Theory (B) 17:48–50

Chang JT (1996a) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134:189–215

Chang JT (1996b) Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Math Biosci 137:51–73

DeBry RW (1992) The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol Biol Evol 9:537–551

Eernisee DJ, Kluge AG (1993) Taxonomic Congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Mol Biol Evol 10:1170–1195

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

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 266:418–427

Feng DF, Doolittle RF (1997) Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. J Mol Evol 44:361–370

Caspers G-J, Reinders G-J, Leunissen JAM, Wattel J, de Jong WW (1996) Protein sequences indicate that turtles branched off from the amniote tree after mammals. J Mol Evol 42:580–586

Grishin NV (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 41:675–679

Grishin NV (1997) Estimation of evolutionary distances from protein spatial structures. J Mol Evol 45:359–369

Gu X, Fu Y-X, Li W-H (1995) Maximum likelihood estimation of the heterogeneity of substitution rates among nucleotide sites. Mol Biol Evol 12:546–557

Hedges SB (1994) Molecular evidence for the origin of birds. Proc Natl Acad Sci USA 91:2621–2624

Hendy MD, Penny D, Steel MA (1994) A discrete Fourier analysis for evolutionary trees. Proc Natl Acad Sci USA 91:3339–3343

Holmquist R, Goodman M, Conroy T, Czelusniak J (1983) The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19:437–448

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed.) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2: 87–90

Kumar S, Rzhetsky A (1996) Evolutionary relationship of eukarytic kingdoms. J Mol Evol 42:183–193

Leblanc C, Boyen C, Richard O, Bonnard G, Grienenberger J-M, Kloareg B (1995) Complete sequence of the mitochondrial DNA of the rhodophyte *Chondrus crispus* (*Gigartinales*). Gene content and genome organization. J Mol Biol 250:484–495

Lee MSY (1997) Reptile relationships turn turtle. Nature 389:245–246

Olsen GJ (1987) Earliest phylogenetic branching: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol 52:825–837

Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 38:642–643

Platz JE, Conlon JM (1997) . . . and turn back again. Nature 389:246

Rieppel O, deBraga M (1994) Turtles as diapsid reptiles. Nature 384: 453–455

Rzhetsky A, Sitnikova T (1996) When is it safe to use an oversimplified substitution model in tree-making. Mol Biol Evol 13:1255–1265

Saitou N (1996) Reconstruction of gene trees from sequence data. Methods Enzymol 266:427–449

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schadt EE, Sinsheimer JS, Lange K (1998) Computational advances in maximum likelihood methods for molecular phylogeny. Genome Res 8:222–233

Steel MA, Szekely LA, Hendy MD (1994) Reconstructing trees when sequence sites evolve at variable rates. J Comput Biol 1:153–163

Stuart A, Ord JK (1994) Kendall's advanced theory of statistics, Vol 1, 6th ed. Halsted Press, John Wiley & Sons, New York, Toronto, pp 350–351

Tajima F, Takezaki N (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol Biol Evol 11: 278–286

Tateno Y, Takezaki N, Nei M (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining and maximum parsimony methods when substitution rate varies with site. Mol Biol Evol 11:261–277

Tourasse NJ, Gouy M (1997) Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol Biol Evol 14:287–298

Uzzell T, Corbin KW (1971) Fitting discrete probability distribution to evolutionary events. Science 172:1089–1096

Wilkinson M, Thorley J, Benton MJ (1997) Uncertain turtle relationship. Nature 387:466

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z (1996a) Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol 42:294–307

Yang Z (1996b) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316–324

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166