Freeing Phylogenies from Artifacts of Alignment[1]

Jeffrey L. Thorne[2] *and Hirohisa Kishino*[3]
Department of Genetics, University of Washington

Widely used methods for phylogenetic inference, both those that require and those that produce alignments, share certain weaknesses. These weaknesses are discussed, and a method that lacks them is introduced. For each pair of sequences in the data set, the method utilizes both insertion-deletion and amino acid replacement information to estimate a pairwise evolutionary distance. It is also possible to allow regional heterogeneity of replacement rates. Because a likelihood framework is adopted, the standard deviation of each pairwise distance can be estimated. The distance matrix and standard error estimates are used to infer a phylogenetic tree. As an example, this method is used on 10 widely diverged sequences of the second largest RNA polymerase subunit. A pseudo-bootstrap technique is devised to assess the validity of the inferred phylogenetic tree.

Introduction

DNA and protein sequences contain valuable phylogenetic information. A plethora of methods for extracting this information exist. Most require sequence alignment and do not utilize insertion-deletion information effectively. Detailed introductions to techniques for phylogenetic reconstruction from aligned sequences can be found in reports by Felsenstein (1988) and Swofford and Olsen (1990). Unfortunately, the correct alignment for a set of DNA or protein sequences is usually unknown. The tree topology inferred by alignment-requiring methods is dependent on the quality of the alignment. When several insertions and deletions have occurred, techniques that require an alignment are subject to error; a flawed alignment can lead to flawed conclusions about evolutionary history. Lake (1991) has described a case where the most strongly supported tree topology depends critically on the alignment.

With the exception of Sankoff and his colleagues (e.g., see Sankoff et al. 1973, 1976; Sankoff 1975), the relationship between alignments and phylogenies has been ignored until recently. In the past decade, several methods have been introduced that simultaneously infer the phylogeny and an alignment that reflects the phylogeny (e.g., see Feng and Doolittle 1987; Konings et al. 1987; Higgins and Sharp 1989; Hein 1990). Because an alignment is designed to represent the evolutionary correspondence between sequences, it is desirable for alignment inference to reflect the fact that sequences are related via an evolutionary tree.

1. Key words: phylogenetic inference, RNA polymerase, maximum likelihood, distance matrix, bootstrap.

2. Present address: Biometrics Unit—337 Warren Hall, Cornell University, Ithaca, New York 14853.

3. Present address: Ocean Research Institute, University of Tokyo, 1-15-1, Minami-dai, Nakano-ku, Tokyo 164, Japan.

Address for correspondence and reprints: Jeffrey L. Thorne, Biometrics Unit—337 Warren Hall, Cornell University, Ithaca, New York, 14853.

Mol. Biol. Evol. 9(6):1148–1162. 1992.

0737-4038/92/0906-0010$02.00

Although alignment-producing methods are superior in some respects to alignment-requiring methods, all alignment-producing methods share several flaws. For example, it is difficult to determine the statistical validity of tree topologies inferred by alignment-producing approaches. For methods that require aligned sequences, the bootstrap resampling procedure (Felsenstein 1985) is available, but it has not been successfully extended to procedures that simultaneously align sequences and reconstruct the phylogeny. Alignment-producing methods create alignments with a built-in phylogenetic structure. These alignments should not be used as the input for alignment-requiring methods. An alignment inferred by aligning according to a specific tree topology will be biased in favor of that topology. When the bootstrap is applied to such an alignment, the result can be an artificially high level of confidence in the tree topology that produced the alignment.

As an illustration, consider four distantly related DNA sequences of the gene for the ribulose biphosphate carboxylase large subunit (*rbcL*). The alignment-producing method implemented in the program Treealign (Hein 1990) allows the user to specify which tree topology will guide alignment inference. This program was used to produce three alignments between the four *rbcL* sequences; each alignment was guided by one of the three possible unrooted topologies.

The treatment of gaps is one problem that can arise when alignment-requiring methods are used. In this analysis, gaps were treated as missing data. In other words, a gap position was treated as a nucleotide of unknown type. Each alignment was analyzed by a parsimony and a maximum-likelihood alignment-requiring method. Bootstrap analyses were performed to assess the support for the phylogenetic inference. The phylogenetic structure imposed by the alignment procedure affected inferences made by the alignment-requiring methods. The results of the alignment-requiring analyses were dependent on the alignment (table 1). The bootstrap frequency of a

Table 1
Phylogenetic Bias That Alignments Can Contain

Topology Used to Guide Alignment	Most Parsimonious No. of Substitutions (bootstrap frequency[a])/ Maximum Log Likelihood (bootstrap frequency[a]), for Indicated Topology		
	[(Alcal, Chrom), (Crypt, Nicot)]	[(Alcal, Nicot), (Chrom, Crypt)]	[(Alcal, Crypt), (Chrom, Nicot)]
[(Alcal, Chrom), (Crypt, Nicot)]	1,309 (0.922)/ −6,452.3 (0.596)	1,401 (0.000)/ −6,499.5 (0.000)	1,328 (0.078)/ −6,456.9 (0.404)
[(Alcal, Nicot), (Chroma, Crypt)]	1,327 (0.863)/ −6,503.6 (0.040)	1,381 (0.000)/ −6,515.3 (0.000)	1,341 (0.137)/ −6,480.9 (0.960)
[(Alcal, Crypt), (Chroma, Nicot)]	1,341 (0.104)/ −6,502.4 (0.001)	1,422 (0.000)/ −6,522.7 (0.000)	1,323 (0.896)/ −6,429.8 (0.999)

NOTE.—The program Treealign was used to infer three alignments between four *rbcL* sequences. The sequences Alcal, Chrom, Crypt, and Nicot are from *Nicotiana tabacum* (Shinozaki and Sugiura 1982), *Alcaligenes eutrophus* (Andersen and Caton 1987), *Cryptomonas phi* (Douglas et al. 1990), and *Chromatium vinosum* (Viale et al. 1989), respectively. Each alignment was produced by using one of the three possible unrooted topologies as a guide. The default Treealign penalties were used. The penalties are as follows: 2 for transitions, 5 for transversions, and $8 + 3k$ for a gap of length k nucleotides. Version 3.4 of the PHYLIP computer package was used to perform both a parsimony analysis and a likelihood analysis of the three alignments. The rows of the table contain results from the analysis of a specific alignment.

[a] Calculated by analyzing 1,000 resampled data sets.

topology is influenced by whether that topology was used to guide the alignment. The magnitude of the alignment effect is large; it reveals the weakness of aligning according to a tree and then using an alignment-requiring method.

These four *rbcL* sequences were selected because they are highly diverged. If these sequences had been more closely related, the effect of alignments on topology inference would probably have been smaller. It is not clear how closely related a group of sequences must be to negate the effect of alignment artifacts on phylogeny inference.

Also, the Treealign program was not designed to be—and has not been described by its author as—a tool for producing an alignment that can be analyzed by alignment-requiring methods. It can simultaneously align sequences and infer their phylogeny but cannot provide confidence statements about the phylogeny. The output alignments from this and similar programs have been misused by researchers who desire both estimates of the phylogeny and bootstrap frequencies. Treealign is not alone in producing phylogenetically biased alignments. Other programs that simultaneously align sequences and infer phylogenies also do so.

Another weakness of alignment-producing methods involves the weighting of gaps in the alignment, versus the weighting of substitutions. Different relative weights will affect alignment inference and, presumably, phylogenetic reconstruction. It is difficult to choose appropriate weights. Widely used implementations of alignment-producing methods use weights chosen arbitrarily rather than objectively; the relative weights are fixed rather than adapted to the data being analyzed.

In our opinion, the greatest weakness of methods that simultaneously align sequences and reconstruct the phylogeny is shared with methods that require the alignment to be given. Both classes of methods use only a single alignment for phylogenetic reconstruction. Ideally, all alignments possible for a set of sequences would contribute to a phylogenetic inference. It is important to consider alternative alignments, because the best single alignment may be misleading. For example, the most probable single alignment between distantly related sequences typically contains fewer gaps than the true alignment (fig. 1).

In a report by Thorne et al. (1991), an evolutionary process that allows insertions and deletions as well as substitutions was studied. A method was developed that allows all possible alignments between a pair of sequences to contribute to evolutionary parameter estimates. This method was compared with one that allows only a single maximum-likelihood alignment to contribute to estimates. For the method that considers only a single alignment, estimates of substitution amounts were positively biased, and estimates of indel amounts were negatively biased. This was not true for the former method; it was more accurate and less biased. For pairs of highly diverged sequences, the superiority of the former method was especially obvious. This implies that phylogenetic inferences based on a single multiple-sequence alignment of a set of highly diverged sequences might also be severely biased. This bias may lead to an error in the inference of evolutionary tree topology.

In the present report, we present a distance-based method for phylogenetic inference, a method that is immune to most of the alignment-induced problems of phylogenetic inference. Our method is presented in terms of protein sequences, but it is easily modified to DNA sequence analysis. It involves obtaining a maximum-likelihood estimate of evolutionary distance between each pair of sequences. All possible pairwise alignments make a contribution to an evolutionary distance estimate. After the distance estimates are found, the phylogeny is inferred by minimizing a weighted

```
A:

ACCGGTACCTGATGCCG-ATTC--GTCTCATGCTT-----T---ATTG-ACTAGAACGTGAACATAACT
C--GGGCCCAGTTACCGAAACACTCTGCAATGCCCACGTTGGCTACGGAAGAAGAC-------CTTCCA

TGAGACCAGACCT--CTATGTAGGTTCAGGCTAATACTTCCTCCGGGAAGACAAAATACGAACCCGGTC
TAGTACCGGCCATTTCATTCGAGTTCGAGGTTTATAGTACCCCTGTTAGGCCCAACTATG----CGGTG

CATGAGTGATAGAACAAACA----------CTACAAC-CTGGTGGTGTAGCTAGGCTCGATTACTACGT
CT--TGGCATATATCAAGCAAGACCGGGCGTGACGAAGCTGGTAGTGTAGTTAGGCACTATA-----AG

GGAGCAAGATGGTCTATTTGTGGTCTTGGTTTCTGAGACTGCATCGACAATTATGATCCTTATCTAGAA
GAGAGAACATGGGCAAGTTGCCGTGTCG--TTCCCAGTTTAGGTCGAAAAGTGGGATGGCATTCGATA-

GCCAATCGTTGGGGGGTGTCAACTCGGCA-TAGCATGTC----------GGCTTCCAG
------AGTCA------TAAA----TCATCTTGACGGAATCCTAGAGTGGGGTCCAC


B:

ACCGGTACCTGATGCCG----ATTCGTCTCATGCTTTATTGACTA-GAACGTGAACATAACTTGAGACCA
--CGGGCCCAGTTACCGAAACACTCTGCAATGCCCACGTTGGCTACGGAAGAAGACCTTCCATAGTACCG

G--ACCTCTATGTAGGTTCAGGCTAATACTTCCTCCGGGAAGACAAAATACGAACCCGGTCCATGAGTGA
GCCATTTCATTCGAGTTCGAGGTTTATAGTACCCCTGTTAGGCCCAACTATG----CGGTGCTT--GGCA

TAGAACAAACACTAC----------AACCTGGTGGTGTAGCTAGGCTCGATTACTACGTGGAGCAAGAT
TATATCAAGCAAGACCGGGCGTGACGAAGCTGGTAGTGTAGTTAGGCACTATAA-----GGAGAGAACAT

GGTCTATTTGTGGTCTTGGTTTCTGAGACTGCATCGACAATTATGATCCTTATCTAGAAGCCAATCGTTG
GGGCAAGTTGCCGTGTCG--TTCCCAGTTTAGGTCGAAAAGTGGGATGGCATTCGATAAGTCATAAATCA

GGGGTGTCAACTCGGCATAGCATGTCGGCTTCCAG
------TCTTGACGGAATCCTAGAGTGGGGTCCAC
```

FIG. 1.—Comparison between a true alignment and an optimal alignment. The insertion-deletion model of Thorne et al. (1992) and the substitution model of Jukes and Cantor (1969) were used to evolve a descendant DNA sequence from an ancestral sequence. The expected number of nucleotide substitutions per alignment position was 0.75. The insertion-deletion parameters (defined in the subsection entitled "The Insertion-Deletion Model") were $r = 0.67$ and $\mu t = 0.1$. The simulation was designed so that the ancestral sequence would have an expected length of 270 nucleotides. A, True alignment between the ancestral and descendant sequences. B, Maximum-likelihood alignment between the ancestral and descendant sequences.

least-squares criterion. To assess the accuracy of the inference, we present a pseudo-bootstrap technique.

The pairwise nature of our method is a limitation. It cannot extract all of the evolutionary information in a multiple-sequence alignment. A more powerful method would allow all or at least a large representative sample of the possible multiple-sequence alignments to contribute to phylogenetic inference. Until such a method is computationally feasible, we believe a conservative approach is warranted when analyzing distantly related sequences. Our approach is conservative; when it does find strong support for an evolutionary hypothesis, one does not need to wonder whether the support is simply an artifact of the method.

The Evolutionary Model

Maximum-likelihood estimation of pairwise distances requires a model of sequence evolution. The evolutionary model presented here allows amino acid replacements as well as insertions and deletions. A weakness of this model is that evolution is assumed to operate directly at the amino acid level and not at the level of the underlying nucleotide code. The probability that one type of amino acid is replaced

by another should depend not only on the two amino acid types but also on the specific nucleotide triplet that codes for the original amino acid. Furthermore, an ideal evolutionary model would not ignore frameshift mutations or the existence and evolution of introns.

Amino Acid Replacement Model

The amino acid replacement process is assumed to be Markovian. Let π_i be the equilibrium probability (i.e., the frequency) of amino acid type i, and let $P_t(j \mid i)$ be the probability that a sequence position occupied by an amino acid of type i is occupied by an amino acid of type j after an amount of evolution equal to t. A reversible amino acid replacement model is a model for which

$$\pi_i P_t(j \mid i) = \pi_j P_t(i \mid j) \tag{1}$$

for all amino acid types i and j. Any reversible amino acid replacement model could be adopted for the estimation of pairwise distances from nonaligned sequences. In the present report, the empirical amino acid transition matrix of Dayhoff et al. (1978) is adopted.

The mathematical aspects of this model have been explained elsewhere (e.g., see Dayhoff et al. 1978; Kishino et al. 1990; but, for criticism of the Dayhoff et al. model, also see Wilbur 1985). To construct the empirical transition matrix, Dayhoff et al. (1978) collected sets of easily aligned (i.e., closely related) sequences. Only closely related sequences were considered, because, when evolutionary distance is sufficiently small, the possibility of multiple replacements can be ignored. The observed replacement patterns were used to construct an amino acid transition matrix.

The Dayhoff et al. model measures evolutionary distance in units called "PAMs" (accepted point mutations per 100 residues). If s is the amino acid replacement rate, then an evolutionary distance of $st = k$ PAMs between two sequences is equivalent to the expectation that the two sequences are separated by k amino acid replacements per 100 residues. Because the model is empirical, it reflects both the fact that different amino acid types are replaced at different rates and the fact that amino acids are usually replaced by biochemically similar amino acids.

The Insertion-Deletion Model

Similar to the Dayhoff et al. model, the insertion-deletion model employed here is Markovian and reversible. It has been presented elsewhere, in the context of DNA sequence evolution (Thorne et al. 1992). Each protein sequence is treated as a sequence of concatenated amino acid fragments. Each fragment contains one or more amino acids. Because fragments are inserted and deleted as units, the fragment-size distribution should be identical to the insertion-deletion-length distribution. For computational convenience, the number of amino acid residues per fragment (i.e., the fragment size) is assumed to be geometrically distributed. At the cost of more computation, other distributional forms could have been adopted. Let $h(n)$ be the probability that a fragment is associated with exactly n amino acids residues. For the geometric distribution, this probability can be expressed as

$$h(n) = (1-r)r^{n-1} \quad 0 \le r \le 1 \quad n = 1,2,\ldots. \tag{2}$$

The expected length of a fragment is $1/(1-r)$.

A sequence that contains n fragments experiences deletions at rate $n\mu$. The same sequence experiences insertions at rate $(n+1)\lambda$, because insertions are allowed only at the $n-1$ interior boundaries between fragments, the extreme 5′ end of the sequence, and the extreme 3′ end of the sequence. The likelihood expression for a pair of sequences involves μt and λt, where t is the divergence time; μ and λ cannot be estimated independently of t. To reduce the number of parameters to be estimated, the value of λt will depend on the value of μt as described in equations (11) and (12) of Thorne et al. (1992). In practice, the effect of this forced dependence is to make λt almost equal to—but slightly less than—μt. The ratio of λt and μt will depend on the lengths of the two sequences being analyzed. As will become clear later, ideally this ratio would be the same for all pairs of sequences in the data set. Fortunately, the variation in this ratio is small enough in biologically interesting data sets to make this imperfection of only trivial importance.

Insertions and deletions at the extreme 5′ or 3′ ends of a sequence must be treated carefully. Terminal gaps in an alignment may not actually be due to insertions or deletions; they could be an artifact of the data collection procedure. Our practice is to consciously avoid terminal gaps; conserved regions are used to delimit the sequence portions that are analyzed. The conscious avoidance of terminal gaps injects a subjective element into an otherwise objective method, but, for reasonably long sequences, this subjectivity should not have a significant impact.

The Method

Likelihood of Two Nonaligned Sequences

The likelihood of two nonaligned sequences depends on the likelihoods of the alignments that can relate them. Each pairwise alignment is a specific hypothesis about the evolutionary relationship between two sequences. Some relationships are more probable than others. The probability of two sequences, given specific parameter values (i.e., the likelihood), is the sum, over all possible relationships (i.e., alignments), of the probabilities of the relationships. The parameter values that maximize the probability of two sequences are the maximum-likelihood estimates. All possible alignments between the sequence pair contribute to these estimates. These estimates can be computed with the methods of Thorne et al. (1991, 1992).

Distance Estimation

For each pair of sequences in a data set, distance-based approaches to phylogenetic reconstruction require an estimate of d_{ij}, the distance between the ith and jth sequences. Under a molecular clock, this distance should have the form $d_{ij} = \alpha t_{ij}$, where α is a constant and t_{ij} is the time since divergence of the ith and jth sequences. The existence of a molecular clock is not crucial to the following discussion; it is adopted only for the sake of explanation. Another measure of branch length could be used in place of divergence time. The methods of Thorne et al. (1991, 1992) are easily adapted to find $\hat{d}_{ij}$, a maximum-likelihood estimate of d_{ij}, from a pair of sequences that are not aligned. Because all possible alignments between the two sequences contribute to the likelihood, these estimates of d_{ij} should be more accurate and less biased than estimates produced from the same data by other approaches.

The amounts of amino acid replacement (st_{ij}), insertion (λt_{ij}), and deletion (μt_{ij}) contribute to the distance between two sequences. Although evolutionary rates are allowed to differ among lineages, we assume here that both the distribution of indel lengths and the ratio of indel rates to amino acid replacement rates are constant among

lineages. In other words, we assume that r and $\rho = s/\mu$ are constant among lineages. Under this assumption, we could adopt any one of st_{ij}, λt_{ij}, and μt_{ij} as a measure of the distance. Certainly, the assumption will sometimes be incorrect, and it is not clear how robust the proposed distance method will be to violations of this assumption.

Assume that we are interested in N sequences. Furthermore, let A_i and A_j be the ith and jth sequences. We could estimate d_{ij}, ρ, and r by maximizing the following pseudo-likelihood function:

$$L = \Pi_{i<j} L(d_{ij}, \rho, r | A_i, A_j) . \quad (3)$$

Because this maximization is computationally prohibitive, we instead adopt a two-step procedure. Our previously published methods are designed to jointly estimate st_{ij}, μt_{ij}, and r from a pair of sequences. Let these maximum-likelihood estimates be $\widehat{st}_{ij}$, $\widehat{\mu t}_{ij}$, and r_{ij}. Also, let $\rho_{ij} = \widehat{st}_{ij}/\widehat{\mu t}_{ij}$. The first step of our procedure is to calculate ρ_{ij} and r_{ij} for each of the $N(N-1)/2$ sequence pairs. Because ρ and r are assumed to be constant, we set ρ equal to the median of the ρ_{ij} values and set r equal to the median of the r_{ij} values. These are the values of ρ and r that will be used for the calculation of pairwise distances.

To estimate ρ and r, we select the medians instead of the means, because medians are more robust to outliers. When a pair of sequences is closely related, ρ and r estimates based solely on this single sequence pair can be poor, because of paucity of evolutionary events. When a pair of sequences is distantly related, ρ and r estimates based solely on the single sequence pair can be poor, because of the difficulty of differentiating between replacement events and insertion or deletion events. A single poor estimate of ρ, for example, could substantially affect the mean of the $N(N-1)/2$ initial estimates while only slightly affecting the median.

After these medians are found, each pair of sequences can be analyzed again. This time, st_{ij} can be considered the only free parameter, because ρ and r are prespecified. The maximum-likelihood estimates of st_{ij} that result from this reanalysis will serve as the pairwise distances.

Enhanced accuracy of distance estimates is one advantage of this maximum-likelihood approach. Another advantage is the ability to approximate the variance of each distance estimate. This approximation can be made by examining the curvature of the log-likelihood surface (e.g., see Kendall and Stuart 1973, pp. 45–46). This feature can prevent a poorly estimated distance from having undue influence on the phylogeny. In contrast, widely used distance methods [e.g., the method of Cavalli-Sforza and Edwards (1967), the method of Fitch and Margoliash (1967), and the neighbor-joining method of Saitou and Nei (1987)] are more susceptible to the influence of a poorly estimated distance on the phylogeny.

After estimation of distances, the proposed method is straightforward. Like both the method of Fitch and Margoliash (1967) and the method of Cavalli-Sforza and Edwards (1967), this method consists of minimizing a weighted least-squares criterion. It can be implemented via a slight modification of FITCH—the least-squares distance program in PHYLIP (Felsenstein 1989). Let z_{ij}^2 be the approximate variance of $\hat{d}_{ij}$. Also, let p_{ij} be the length of the path that connects the ith and jth sequences on a tree, and let

$$S = \sum_{i<j} \frac{(\hat{d}_{ij} - p_{ij})^2}{z_{ij}^2}. \quad (4)$$

The proposed method selects the tree that minimizes S.

Accuracy of the Inferred Topology: Pseudo-Bootstrap

The value of this method would be enhanced if the support for the inferred phylogeny could be assessed. Although pairwise distance estimates are not independent, one possibility is to treat them as if they were and then show by simulation that this treatment is reasonable. The idea is to generate, via a parametric bootstrap, "new" sets of distances from the original set. Each new set can be used to reconstruct the phylogeny. The variability of the phylogenies that are reconstructed from new sets of distances can be used to assess the support for the original inference.

If sequences i and j are long, then the distribution of $\hat{d}_{ij}$ will resemble a normal distribution with mean d_{ij} and variance z_{ij}^2 (e.g., see Kendall and Stuart 1973, pp. 45–46). To assess the accuracy of the inferred phylogeny, we sampled a new distance d_{ij}^* for each i and j from a normal distribution with mean $\hat{d}_{ij}$ and variance z_{ij}^2. A phylogeny can be inferred from this simulated data set by minimizing S^*, where

$$S^* = \sum_{i<j} \frac{(d_{ij}^* - p_{ij}^*)^2}{z_{ij}^2} \quad (5)$$

and where p_{ij}^* is the length of the path that connects the ith and jth sequences on this tree. This process of creating resampled distance matrices and then inferring a phylogeny can be repeated many times. As explained by Felsenstein (1985), consensus trees can be constructed from these inferred phylogenies and can be used to assess the support for the original inference.

This resampling approach incorrectly treats pairwise distances as if they were independent. To correctly create resampled distance matrices, the correlations between pairwise distances should be considered (Hasegawa et al. 1985). Phylogenetic-support statements produced by this approach will tend to be conservative. For example, this approach might yield a pseudo-bootstrap frequency of 90% when a better approach that correctly accounts for pairwise correlations would yield a confidence level of 95%. The conservative nature of this approach stems from the treatment of pairwise distances as being independent. Pairwise distance correlations are imposed by the phylogenetic structure. The correlations exist because the path that connects two tips of a tree often will traverse some of the same branches that are traversed by a path connecting another pair of tips; random evolutionary events that occur on the common branches will affect both pairwise distances in the same direction. A resampled distance matrix created by ignoring these correlations will contain less phylogenetic structure. This is revealed in practice by the fact that S^*, the criterion minimized when a phylogeny is inferred from a resampled distance matrix, is invariably greater than S, the criterion minimized when a phylogeny is inferred from the actual distance matrix.

Differences between support levels produced by this pseudo-bootstrap approach and those produced by the conventional bootstrap of prealigned sequences were explored through simulation. To assist this simulation study, two computer programs were provided by Joseph Felsenstein. One of these generated random evolutionary trees via a branching process, and the other evolved DNA sequences of length 250

nucleotides along the trees. Trees were generated by successive splittings of evolutionary lineages. The first splitting event defined the root of the tree. Subsequent events created interior nodes of the tree. Trees with 10 tips were simulated; simulations were stopped immediately before occurrence of the split that would create the 11th tip. If termination of the simulations had not occurred, then individual branch lengths on these trees would have been exponentially distributed with a mean of 0.1 substitutions per sequence position. Because termination did occur, the actual distribution was somewhat different.

Because the conventional bootstrap requires an alignment, the simulated evolutionary process involved only substitutions. Specifically, the substitution model of Jukes and Cantor (1969) was employed. Although rooted topologies were generated, the proposed distance method was used to infer unrooted topologies. Values of z were computed by the formula of Kimura and Ohta (1972). Each interior branch of a topology serves to partition the sequences into two groups, and the bootstrap support of these partitions or interior branches can be measured. In total, 10 randomly generated topologies were studied. For each interior branch, a conventional bootstrap frequency was computed by analyzing 1,000 resampled data sets, and a pseudo-bootstrap frequency was computed by analyzing 1,000 resampled distance matrices.

Of the 70 interior branches defined by the 10 random topologies, 50 had a conventional bootstrap frequency of ≥ 0.9. These 50 conventional bootstrap frequencies were placed into four categories: 0.9–0.949, 0.95–0.989, 0.99–0.999, and 1.0. There were no cases of a pseudo-bootstrap frequency being >0.9 when the conventional bootstrap frequency was <0.9. Table 2 describes the distribution of pseudo-bootstrap frequencies among all conventional bootstrap frequencies that fall into a given interval. Clearly, the pseudo-bootstrap frequencies are conservative. The difference between the two types of bootstrap frequencies appears to grow as the conventional bootstrap frequency decreases. Only bootstrap frequencies corresponding to interior branches of the true tree are summarized in table 2. There was one case of a partition that was not on the true tree and that had a conventional bootstrap frequency >0.9 (0.932). It is interesting that the pseudo-bootstrap frequency of this partition was <0.9 (0.696).

This was only a preliminary comparison of the conventional bootstrap and the pseudo-bootstrap. Although a more general simulation study is warranted, this preliminary investigation is consistent with the expected conservative nature of the pseudo-bootstrap, and it hints that the pseudo-bootstrap is not so conservative as to be without value. A more accurate procedure for assessing the support of inferences made by the proposed distance method would be desirable, but this pseudo-bootstrap procedure errs on the side of caution.

Table 2
Comparison of Conventional Bootstrap and Pseudo-Bootstrap

Conventional Bootstrap Frequency	No. of Instances with Pseudo-Bootstrap Frequency of						
	1.000	0.990–0.999	0.950–0.989	0.900–0.949	0.800–0.899	0.700–0.799	0.600–0.699
1.000	17	6	1	0	0	0	0
0.990–0.999	2	4	2	1	1	0	0
0.950–0.989	0	0	2	4	1	2	0
0.900–0.949	0	0	0	0	5	1	1

An Example

To demonstrate this method, a phylogeny is inferred from 10 amino acid sequences of the second largest RNA polymerase subunit. These sequences include two eukaryotic pol I sequences, two eukaryotic pol II sequences, two eukaryotic pol III sequences, two archaebacterial sequences, a eubacterial sequence, and a chloroplast sequence. The 10 sequences possess a relatively conserved region near their 5′ end and near their 3′ end. Only the amino acids between and including these two conserved regions were considered. The footnote to table 3 contains names and sources of sequences, as well as specification of where conserved regions begin and end.

A phylogenetic analysis of these 10 sequences was performed. The first step was to find ρ and r from the medians of the 45 possible pairwise comparisons between the RNA polymerase sequences. The resulting values of ρ and r were, respectively, 29 and 0.89. With these values, pairwise distances and approximate standard deviations were computed (table 3). The distances were used to infer a phylogeny via the proposed method (fig. 2).

The amount of computation required by this method is large but not prohibitive. First-step pairwise comparisons (i.e., those done to find ρ and r) are slower than second-step pairwise comparisons (i.e., those done to find pairwise distances), because first-step comparisons require that three parameters (μt_{ij}, st_{ij}, and r) be estimated instead of just one (d_{ij}). For example, the first-step comparison between the *Drosophila* pol I and pol II sequences required 54.8 min of central processing unit (CPU) time, and the second-step comparison required 12.5 min of CPU time, on a Sun Sparcstation IPX.

The Fitch-Margoliash and neighbor-joining implementations in PHYLIP version 3.4 were also applied to this distance matrix. The option that does not allow the Fitch-

Table 3
Pairwise Distances and Standard Errors

	Sc1	Dr1	Sc2	Dr2	Sc3	Dr3	Sul	Met	Esc	Spi
Sc1		4.3	6.6	6.9	7.0	6.5	7.0	6.3	10.1	10.3
Dr1	81.9		6.6	7.2	7.1	7.4	7.0	6.8	10.2	10.5
Sc2	112.3	112.3		2.6	4.4	4.9	4.3	4.3	8.7	9.3
Dr2	113.6	118.7	48.1		4.8	5.0	4.3	4.3	8.5	9.0
Sc3	108.2	122.8	86.0	90.2		2.8	4.9	4.9	8.7	9.5
Dr3	108.7	124.9	92.0	95.9	49.2		5.3	5.2	8.7	9.4
Sul	117.2	117.9	77.3	80.7	91.7	100.6		3.2	6.8	8.8
Met	111.5	116.9	87.6	86.1	92.5	94.5	61.2		6.5	8.3
Esc	152.2	151.4	136.9	137.1	138.4	140.4	112.0	115.1		4.6
Spi	153.7	162.9	142.4	139.9	144.9	142.6	134.9	134.4	88.2	

NOTE.—The name, abbreviation used, reference, beginning hexanucleotide, ending hexanucleotide, and length of the analyzed sequences are as follows: *Saccharomyces cerevisiae* pol I, Sc1, Yano and Nomura (1991), HIGSFN, EKIFED, 1,110 amino acids; *Drosophila melanogaster* pol I, Dr1, Kontermann et al. (1989), HVDSFD, ARFKLN, 1,089 amino acids; *S. cerevisiae* pol II, Sc2, Sweetser et al. (1987), QLDSFN, PRLYTD, 1,173 amino acids; *D. melanogaster* pol II, Dr2, Falkenburg et al. (1987), QLDSFD, PRLMVT, 1,133 amino acids; *S. cerevisiae* pol III, Sc3, James et al. (1991), HLDSFN, PRLRLE, 1,084 amino acids; *D. melanogaster* pol III, Dr3, Seifarth et al. (1991), HIDSFN, PKMILE, 1,080 amino acids; *Sulfolobus acidocaldarius* RNA polymerase, Sul, Pühler et al. (1989), HLDSFN, PRLILG, 1,092 amino acids; *Methanobacterium thermoautotrophicum* RNA polymerase, Met, Berghöfer et al. (1988), HIHSYN, PKLVLE, 1,095 amino acids; *Escherichia coli* RNA polymerase, Esc, Ovchinnikov et al. (1982), QLDSFQ, INIELE, 1,310 amino acids; and *Spinacia oleracea* chloroplast RNA polymerase, Spi, Hudson et al. (1988), QFEGFW, LNHFLV, 1,040 amino acids. Because the *Methanobacterium* sequence is split into two subunits, the analyzed portion is actually a concatenation of the 3′ end of the B″ subunit sequence and the 5′ end of the B′ subunit sequence. Data below the diagonal are pairwise distance estimates and data above the diagonal are standard errors.

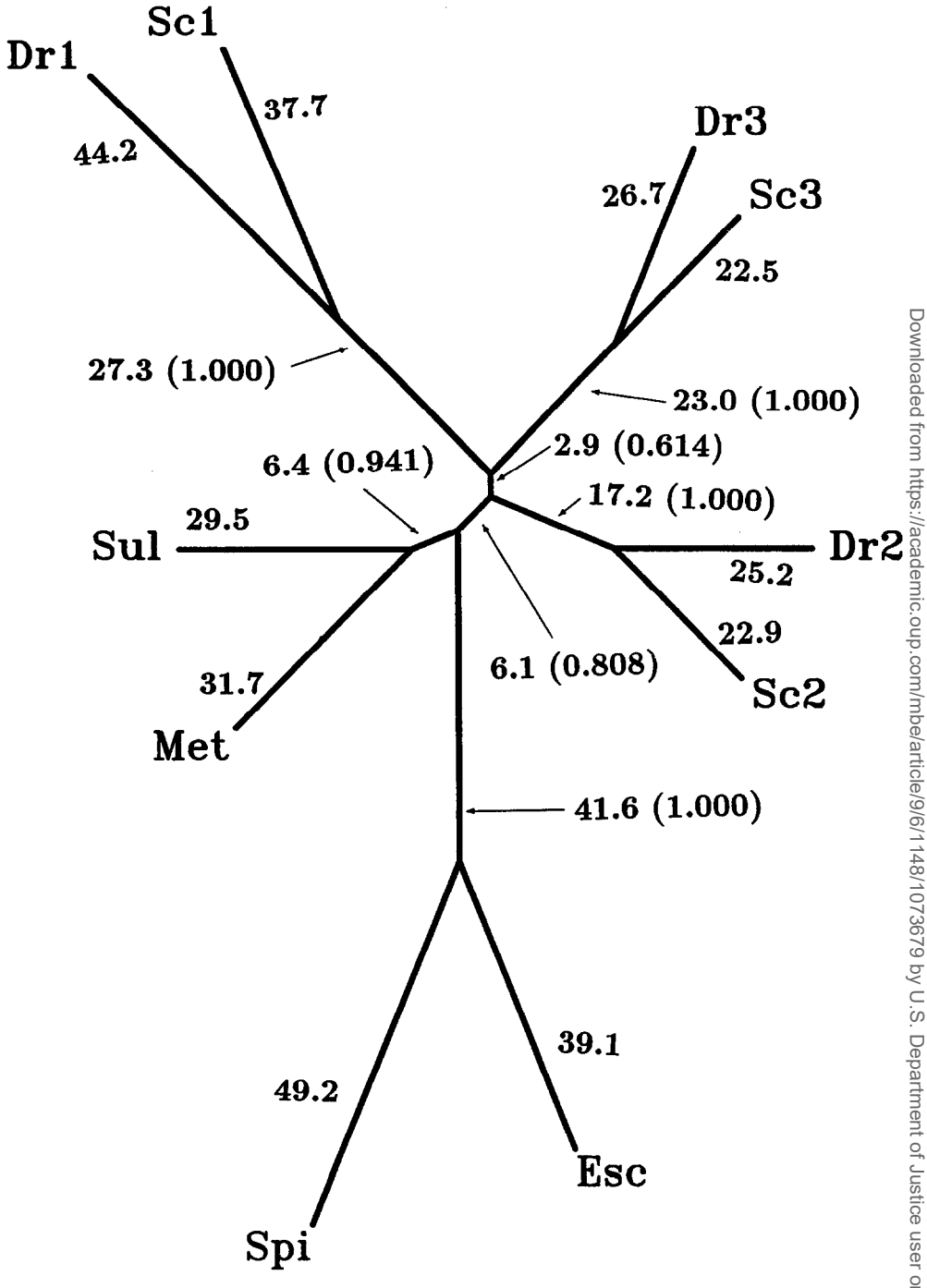

FIG. 2.—Inferred RNA polymerase phylogeny and pseudo-bootstrap frequencies. Abbreviations for species names are as in the footnote of table 3. To obtain the pseudo-bootstrap frequencies, the distance matrix was resampled 1,000 times. A tree was inferred from each of the resampled distance matrices, and these 1,000 inferences were summarized by the program Consense in version 3.4 of the PHYLIP package. The pseudo-bootstrap frequencies of the partitions, defined by the interior branches, are in parentheses.

Margoliash method to infer negative branch lengths was chosen. The Fitch-Margoliash method finds a tree with the same topology as shown in figure 2. The topology found by the neighbor-joining method is slightly different. Whereas the tree in figure 2 contains a branch that separates the pol I and pol III lineages from all other lineages, the neighbor-joining tree instead contains a branch that separates the pol I and pol II lineages from all others. Because the pseudo-bootstrap frequency of the branch that separates the pol I and pol III lineages from all others is low (0.614), we know little about the relative merits of the two topologies.

For the sake of simplicity, we have described our method in the context of regional homogeneity of amino acid replacement rates. In fact, the regions of a protein do not all evolve at the same rate. The evolutionary model of Thorne et al. (1992) can allow regional heterogeneity of replacement rates. Our method is easily modified to correspond to this model. For the polymerase sequences, there is a large difference between the log likelihood produced by a homogeneity analysis and that produced by a heterogeneity analysis. This is an indication that the heterogeneity analysis is more appropriate. For example, the difference is 23.2 log-likelihood units for the first-step comparison between the *Drosophila* pol I and pol II sequences. In light of the fact that the regional heterogeneity model contains only two more parameters than does the homogeneity model, this difference is large. When our method is modified to permit regional heterogeneity, the topology of figure 2 is again preferred.

Discussion

Blaisdell (1986, 1991) has suggested an alignment-free technique for computing pairwise distances between sequences. This approach involves counting, within each sequence, the number of occurrences of all subsequences of length n. A subsequence of length n is called an n-tuple. For example, there are four possible 1-tuples (e.g., A, G, T, and C) and 16 possible 2-tuples in a DNA sequence. To calculate a distance, the number of occurrences of a specific n-tuple in one sequence is subtracted from the number of occurrences in the other, and then this difference is squared. The sum of this difference taken over all n-tuples is used as the pairwise distance. This is a clever approach to avoiding alignment artifacts, but it possesses certain weaknesses. The values of distances calculated in this way depend on both the lengths of the sequences and the amount of evolution that separates them. The choice of tuple size is important, and the n that is the best choice for one pair of sequences may not be best for another pair of sequences. Most important, the Blaisdell method discards some evolutionary information that could be extracted from a sequence pair.

Our method utilizes more evolutionary information than does the Blaisdell method, but it is still incomplete. The pseudo-bootstrap procedure and the treatment of terminal indels both need to be improved. Frameshifts, intron evolution, and complex sequence rearrangements should not be ignored. Also, by considering only pairwise information, this method shares the weaknesses of other distance methods.

While the proposed method is not ideal, it has advantages over widely used phylogenetic inference techniques. Regional heterogeneity of replacement rates can be allowed. The method does not rely on a single alignment or ignore the evolutionary information provided by insertions and deletions. It incorporates the uncertainty of distance estimates into phylogeny estimation. Finally, it has been intentionally designed to be conservative. The pseudo-bootstrap frequency of the inferred topology is an underestimate of statistical support. Alignment-requiring methods may yield artificially high bootstrap frequencies because they ignore uncertainty due to artifacts of alignment.

In the reconstruction of evolutionary history, it is our opinion is that it is better to err on the side of caution.

Acknowledgments

We thank Joseph Felsenstein, Walter Fitch, Linda Hardison, Ben Hall, Mary K. Kuhner, and Bill Hatheway for their help. The research and computing facilities were supported by NSF grant BSR 8918333 and NIH grant 5R01 GM41716 (principal investigator, Joseph Felsenstein). J.L.T. was also supported by a National Science Foundation Graduate Fellowship.

LITERATURE CITED

ANDERSEN, K., and J. CATON. 1987. Sequence analysis of the *Alcaligenes eutrophus* chromosomally encoded ribulose biphosphate carboxylase large and small subunit genes and their gene products. J. Bacteriol. **169**:4547–4558.

BERGHÖFER, B., L. KRÖCKEL, C. KÖRTNER, M. TRUSS, J. SCHALLENBERG, and A. KLEIN. 1988. Relatedness of archaebacterial RNA polymerase core subunits to their eubacterial and eukaryotic equivalents. Nucleic Acids Res. **16**:8113–8128.

BLAISDELL, B. E. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. USA **83**:5155–5159.

———. 1991. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems. J. Mol. Evol. **32**:521–528.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution **21**:550–570.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence structure. Vol. **5**, suppl. 3. National Biomedical Research Foundation, Washington D.C.

DOUGLAS, S. E., D. G. DURNFORD, and C. W. MORDEN. 1990. Nucleotide sequence of the gene for the large subunit of ribulose-1,5-biphosphate carboxylase/oxygenase from cryptomonas phi: evidence supporting the polyphyletic origin of plastids. J. Phycol. **26**:500–508.

FALKENBURG, D., B. DWORNICZAK, D. M. FAUST, and E. K. F. BAUTZ. 1987. RNA polymerase II of *Drosophila:* relation of its 140,000 M_r subunit to the β subunit of *Escherichia coli* RNA polymerase. J. Mol. Biol. **195**:929–937.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1989. Phylip—phylogeny inference package (version 3.2). Cladistics **5**:164–166.

FENG, D.-F., and R. F. DOOLITTLE. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. **25**:351–360.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**: 279–284.

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.

HEIN, J. 1990. A unified approach to alignment and phylogenies. Pp. 626–645 *in* R. F. DOOLITTLE, ed. Methods in enzymology. Vol. **183.** Academic Press, San Diego.

HIGGINS, D. G., and P. M. SHARP. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS **5**:151–153.

HUDSON, G. S., T. A. HOLTON, P. R. WHITEFELD, and W. BOTTOMLEY. 1988. Spinach chloroplast *rpoBC* genes encode three subunits of the chloroplast RNA polymerase. J. Mol. Biol. **200**:639–654.

JAMES, P., S. WHELEN, and B. D. HALL. 1991. The RET1 gene of yeast encodes the second-largest subunit of RNA polymerase III. J. Biol. Chem. **266**:5616–5624.

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

KENDALL, M., and A. STUART. 1973. The advanced theory of statistics, Vol. **2**, 3d ed. Charles Griffen, London.

KIMURA, M., and T. OHTA. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2**:87–90.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

KONINGS, D. A. M., P. HOGEWEG, and B. HESPER. 1987. Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus. Mol. Biol. Evol. **4**:300–314.

KONTERMANN, R., S. SIZLER, W. SEIFARTH, G. PETERSON, and E. K. F. BAUTZ. 1989. Primary structure and functional aspects of the gene coding for the second largest subunit of RNA polymerase III of Drosophila. Mol. Gen. Genet. **219**:373–380.

LAKE, J. A. 1991. The order of sequence alignment can bias the selection of tree topology, Mol. Biol. Evol. **8**:378–385.

OVCHINNIKOV, Y. A., G. S. MONASTYRSKAYA, V. V. GUBANOV, S. O. GURYEV, I. S. SALOMATINA, T. M. SHUVAEAE, V. M. LIPKIN, and E. D. SVERDLOV. 1982. The primary structure of *E. Coli* RNA polymerase: nucleotide sequence of the *rpoC* gene and amino acid sequence of the β′-subunit. Nucleic Acids Res. **10**:4035–4044.

PÜHLER, G., F. LOTTSPEICH, and W. ZILLIG. 1989. Organization and nucleotide sequence of the genes encoding the large subunits A, B, and C of the DNA-dependent RNA polymerase of the archaebacterium *Sulfolobus acidocaldarius.* Nucleic Acids Res. **17**:4517–4534.

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

SANKOFF, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. **78**:35–42.

SANKOFF, D., R. J. CEDERGREN, and G. LAPALME. 1976. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. J. Mol. Evol. **7**:133–149.

SANKOFF, D., C. MOREL, and R. J. CEDERGREN. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. Nature New Biol. **245**:232–234.

SEIFARTH, W., G. PETERSEN, R. KONTERMANN, M. RIVA, J. HUET, and E. K. F. BAUTZ. 1991. Identification of the genes coding for the second-largest subunits of RNA polymerases I and III of *Drosophila melanogaster.* Mol. Gen. Genet. **228**:424–432.

SHINOZAKI, K., and M. SUGIURA. 1982. The nucleotide sequence of the tobacco chloroplast gene for the large subunit of ribulose-1,5-biphosphate carboxylase/oxygenase. Gene **20**:91–102.

SWEETSER, D., M. NONET, and R. A. YOUNG. 1987. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proc. Natl. Acad. Sci. USA **84**:1192–1196.

SWOFFORD, D. L., and G. J. OLSON. 1990. Phylogeny reconstruction. Pp. 411–501 in D. M. HILLIS and C. MORITZ, eds. Molecular systematics. Sinauer, Sunderland, Mass.

THORNE, J. L., H. KISHINO, and J. FELSENSTEIN. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. **33**:114–124.

———. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. **34**:3–16.

VIALE, A. M., H. KOBAYASHI, and T. AKAZAWA. 1989. Expressed genes for plant-type ribulose-1,5-biphosphate carboxylase/oxygenase in the photosynthetic bacterium *Chromatium vinosum,* which possesses two complete sets of the genes. J. Bacteriol. **171**:2391–2400.

WILBUR, W. J. 1985. On the PAM matrix model of protein evolution. Mol. Biol. Evol. **2**:434–447.

YANO, R., and M. NOMURA. 1991. Suppressor analysis of temperature-sensitive mutations of

the largest subunit of RNA polymerase I in *Saccharomyces cerevisiae:* a suppressor gene encodes the second-largest subunit of RNA polymerase I. Mol. Cell. Biol. **11**:754–764.

WALTER M. FITCH, reviewing editor

Received December 27, 1991; revision received June 15, 1992

Accepted June 15, 1992