J Mol Evol (1989) 29:526–537

Journal of
Molecular Evolution

Effectiveness of Measures Requiring and Not Requiring Prior Sequence Alignment for Estimating the Dissimilarity of Natural Sequences

B. Edwin Blaisdell

Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, California 94306, USA

Summary. Various measures of sequence dissimilarity have been evaluated by how well the additive least squares estimation of edges (branch lengths) of an unrooted evolutionary tree fit the observed pairwise dissimilarity measures and by how consistent the trees are for different data sets derived from the same set of sequences. This evaluation provided sensitive discrimination among dissimilarity measures and among possible trees. Dissimilarity measures not requiring prior sequence alignment did about as well as did the traditional mismatch counts requiring prior sequence alignment. Application of Jukes–Cantor correction to singlet mismatch counts worsened the results. Measures not requiring alignment had the advantage of being applicable to sequences too different to be critically alignable. Two different measures of pairwise dissimilarity not requiring alignment have been used: (1) multiplet distribution distance (MDD), the square of the Euclidean distance between vectors of the fractions of base singlets (or doublets, or triplets, or . . .) in the respective sequences, and (2) complements of long words (CLW), the count of bases not occurring in significantly long common words. MDD was applicable to sequences more different than was CLW (noncoding), but the latter often gave better results where both measures were available (coding). MDD results were improved by using longer multiplets and, if the sequences were coding, by using the larger amino acid and codon alphabets rather than the nucleotide alphabet. The additive least squares method could be used to provide a reasonable consensus of different trees for the same set of species (or related genes).

Offprint requests to: B.E. Blaisdell

Key words: Evolutionary trees — Additive least squares — DNA — Greatly divergent sequences — Consensus trees

Introduction

All three well-known conventional methods of inferring evolutionary trees from similarity measures of sequence data, nucleotide base or amino acid, assume prior correct homologous total alignment of the sequences. The biological and rational bases of the three methods (pairwise distances, parsimonious evolution, and compatibility) recently have been reviewed critically by Felsenstein (1982). In addition to the reservations raised by him, the difficult problem of achieving a correct homologous total alignment of the sequences constitutes a significant impediment to a rationally satisfying inference of evolutionary trees from sequence data. The two most commonly used alignment methods are dynamic programming optimization (Needleman and Wunsch 1970; Smith and Waterman 1981) and dot-matrix (Konkel et al. 1979). The Needleman–Wunsch algorithm requires the user to supply values of parameters specifying the penalty for the introduction of a gap into a sequence. No algorithm for a statistically validated choice of the gap penalty has been published. The dot-matrix algorithm requires the user to supply a dot criterion of m matches per n adjacent elements in two sequences. No algorithm for a statistically validated choice of the dot criterion has been published. Furthermore, the method provides no total alignment of stretches between the sequence segments defined by visually selected linear clusters of dots. Because of excessively long run-

ning time proportional to the cube of sequence length for three sequences (Needleman–Wunsch), and because of difficulties in presentation of results (dot-matrix), neither method is applicable practically to alignment of more than two short sequences at a time.

Because of the refractory nature, yet fundamental importance to the usual inference methods, of the alignment problem, I have proposed a sequence dissimilarity measure, MDD, that is not dependent on sequence alignment (Blaisdell 1986). Similar uses of amino acid composition (Cornish-Bowden 1979) and amino acid doublet composition (Gibbs et al. 1971) have been proposed. The objective of this paper is to make an extended systematic test of this suggestion, for α- and β-globin sequences, by making comparisons of many mismatch count measures of sequence dissimilarity requiring prior sequence alignment with the new methods not requiring sequence alignment. It considers the questions of assessment of (1) the validity of the various dissimilarity measures for inferring evolutionary trees (do they produce the same "best" tree?), (2) their power to discriminate the best from alternative trees, (3) the congruence and significance of their estimates of the branch lengths, (4) their power to cope with sequences unalignable because of too great an evolutionary divergence, and (5) the reconciliation of trees for the same species set from different data sets.

Data

The sequence data are taken from GenBank. They comprise eight α-globin coding sequences: Human, Rabbit, Goat, horSe, Mouse, Chicken, Duck, and *Xenopus*; and eight β-globin sequences: the same except Bovine in place of horSe (the uppercase letters are used to identify the species in the tables). Lengthy noncoding sequences are also included when these are available: 200 bases of the 5′ header, 200 bases of the 3′ trailer, and both introns. For the α-globins, noncoding data for six species are available: Human, Goat, horSe, Mouse, Chicken, and Duck; and for β-globins, seven species: Human, Rabbit, Goat, Bovine, Mouse, Chicken, and *Xenopus*. Only one (Human) of the very closely similar primate sequences is used.

Methods

Mismatch counts for DNA coding sequences have been used as the conventional dissimilarity measure requiring prior sequence alignment. I have aligned the coding sequences without gaps as is justified by the extensive data of Dayhoff (1979) on the amino acid sequences of the globins of many species. However, the method here has been extended by using in addition to the original 4-letter base alphabet (T, C, A, G), with and without Jukes–Cantor "improvement," and the familiar 20-letter amino acid alphabet, further alphabets obtained by recoding the sequence (Blaisdell 1983; Karlin et al. 1984). For example, results are given for the 16-letter alphabet of base doublets, the 64-letter alphabet of base triplets, the 256-letter alphabet of base quartets, the 61-letter alphabet of codons, the 3721-letter alphabet of codon doublets, and the 400-letter alphabet of amino acid doublets. The method also has been extended by using, in addition to the usual mismatch count for the entire available sequence, counts for meaningful subsets. Results are given for mismatch counts in the separated exons 1, 2, and 3, the separated codon sites 1, 2, and 3, and for the separated classes of substitutions: transitions (T ↔ C, A ↔ G), transversions preserving weak (strong) base hydrogen-bonding character in the DNA double helix (T ↔ A, C ↔ G), and transversions not preserving these characters (T ↔ G, C ↔ A). For a complete list of the alphabets and sequence subsets used, see Table 4. These many alphabets and sequence subsets have been examined to determine the extent of agreement of the evolutionary trees inferred from them using conventional mismatch counts and the extent of the agreement of these trees with those inferred using the two newly proposed dissimilarity measures. Different data sets should produce the same tree when they are derived from the same set of sequences. Also, the α- and β-globin sequences should yield the same tree, as much evidence shows that the major α- and β-globins diverged long before the divergence of the species considered here (Dickerson and Geis 1983).

Results from two classes of dissimilarity measures not requiring prior sequence alignment have been used. One class, MDD (multiplet distribution distance), measures dissimilarity by the squared Euclidean distance between pairs of vectors of counts of the letters in the chosen alphabets in the chosen subset of bases [e.g., base doublets in the total of coding sequences (Table 4, row 17) or in the total of noncoding sequences (Table 4, row 27)] (Blaisdell 1986). Formally,

$$d_m = \sum_{i=1}^{m} (c_i^a - c_i^b)^2$$

where m is the number of letters in the alphabet and $c_i^a(c_i^b)$ is the count of letter i in sequence a(b).

The second class, CLW (complement of common long words), measures dissimilarity by the total length of sequences not covered by SIBs, SISs, or SABs. A SIB is a significantly long identity block, namely, a word (contiguous sequence of letters) identical in both sequences and longer than the longest word expected in common between two random (independently generated) sequences, each of the same letter composition as that observed. Such long common words are found quickly and easily by an algorithm, linear in total sequence length, described by Karlin et al. (1988a). Significantly long common words are identified as follows (Karlin et al. 1989). Let

$$K_0 = \ln(N_1N_2)/(-\ln \lambda) + z_0$$

where N_1 and N_2 are the numbers of letters in sequences 1 and 2,

$$\lambda = \sum_{i=1}^{m} p_iq_i$$

where m is the number of letters in the alphabet (e.g., 4 for bases and 20 for amino acids) and $p_i(q_i)$ is the fraction of letter i in sequence 1(2) and

$$z_0 = \ln\left[\frac{-\ln(1 - 0.01)}{1 - \lambda}\right]/\ln \lambda$$

Then common words of length $\geq K_0$ have significance levels $P \leq 0.01$.

For example, for the coding sequences of human and rabbit β-globins, $N_1 = N_2 = 441$ and (T, C, A, G) = (0.238, 0.259, 0.195, 0.308), and (0.247, 0.236, 0.211, 0.306), respectively. $\lambda = 0.255$, $z_0 = 3.151$, $K_0 = 12.063$, so that common words of

length ≥13 would occur in less than 1% of such pairs of sequences of random base distribution. In fact, there are common words of lengths 14, 15, 18, 20, 23, 26, 35, and 62, and three of length 17 for (human, rabbit), of lengths 14, 15, and 19 for (human, chicken), but none longer than 9 for (human, *Xenopus*). Thus, for the (human, chicken) pair the CLW distance is (441 − 14 − 15 − 19) = 393.

Similarly, a SIS is a significantly long identity string of identical words of length at least 5, interrupted by mismatch error blocks of length at most 3. The limits of 5 and 3 have been chosen on an ad hoc basis to eliminate too sparsely matching strings from consideration. The first and last elements of mismatch error blocks are mismatches in the two sequences, but the interior elements need not mismatch nor even be of equal number. This latter condition permits deletions (or insertions) of one element. Every SIB is a SIS. Let

$$z_1 = \ln\left[\frac{-\ln(1 - 0.01)\lambda^s n_1! n_2! n_3!}{(1 - \lambda)^{s+1}}\right]/\ln \lambda$$

and

$$K_1 = (\ln(N_1N_2) + n\{\ln[\ln(N_1N_2)] - \ln(-\ln\lambda)\})/(-\ln\lambda) + z_1$$

where n_i is the number of error blocks of length i, $n = \Sigma n_i$ is the total number of interrupting error blocks, and $s = \Sigma i n_i$ is the total number of letters in all the error blocks. Then SISs containing a total of matches $\geq K_1$ have significance levels $P \leq 0.01$.

A SAB is a significant aligned block of length at least 5 that is not statistically significant in isolation but gains significance by being sufficiently close to a SIS (but more distant than the error block length of 3) and by being sufficiently well aligned with it (Karlin et al. 1988b). The expected length K_2 of a SAB depends on two new parameters, L = its distance in letters from the nearest SIS, and d = the difference in L for the two sequences, and is

$$K_2 = 0.5 + \{\ln L + \ln(1 + d) + 0.577 + \ln[\lambda(1 - \lambda)]\}/(-\ln \lambda)$$

and its standard deviation is

$$-\pi/(\sqrt{6} \ln \lambda)$$

Rough values of significance level may be obtained using the normal approximation.

In this report these two measures of dissimilarity have been used only for pairs of sequences. However, both may be used to yield measures of global similarity for arbitrary sets of s sequences. The vector distance method, MDD, then becomes in the usual terminology a test of the homogeneity of s transition matrices of a Markov chain of the appropriate order (Blaisdell 1986). Significantly long common words may be found for any subset of r out of s segments (Karlin et al. 1989). This latter method also provides statistically significant alignments of compatibly aligned significantly long common words and strings (SIBs and SISs), and of shorter words significantly aligned compatibly with these (SABs). The method says nothing about the possibility of compensating gaps in a sequence between SIBs or SISs in compatibly aligned sets, though it may be assumed that they are nonexistent. It also may be assumed that the net gap between neighboring incompatibly aligned sets equals the difference in their displacements from some fiducial location, but how this net gap is to be apportioned among one or more gaps in the several sequences is not apparent. The fiducial location may be any uniquely defined location present in every sequence of the set, for example, the end of the first exon or the beginning of a significantly long common word.

The evolutionary tree is inferred from the pairwise distance measures (dissimilarity measures) by the additive least squares method of Cavalli-Sforza and Edwards (1967). This method generally gave the best result of eight different methods on model trees generated by computer simulation (Astolfi et al. 1981). The method has also been found to be better than parsimony methods by Dayhoff (1979), Saitou and Nei (1987), and Sourdis and Nei (1988). The method was executed using program FITCH in the PHYLIP package (version 2.9) of Felsenstein (1986) in the global (G) nonnegative additive least squares mode ($P = 0.0$), and sometimes with user-supplied trees (mode U). I made small additions to the program to evaluate and print out a dimensionless global measure of the tree's fit to the observed pairwise distance measures, the variance ratio

$$F = \frac{(\Sigma D^2_{obs} - \Sigma R^2)/N_B}{\Sigma R^2/(N_D - N_B)} \quad (1)$$

(variance of observed distances explained by the nonnegative least squares fit)/(variance not explained). Here D_{obs} is an observed pairwise distance measure, R is the residual $D_{obs} - D_{calc}$, D_{calc} by least squares, N_B is the number of branches in the unrooted tree, and N_D is the number of observed pairwise distance measures. For example, for the five-species tree of Table 1, $N_B = (2)(5) - 3 = 7$, $N_D = \binom{5}{2} = 10$. F is the conventional F-statistic for least square fits, but no attempt is made to convert it to significance probabilities, as the assumption of normality and independence of the population of D_{obs} is dubious. Only the montonicity of the values is used: higher values of F correspond to better fits to the data. The tree corresponding to the highest F-value is called the "best" tree.

Because the least squares fitting depends only on the differences among the observed distances (dissimilarity measures), the value of F is inflated by the addition of a constant to all distances, as the sum of squares of residuals is unchanged but the sum of squares of observed distances is increased. For this reason, the program was changed further by norming the raw data, $D_{obs} = m$, prior to further treatment, to m* of a constant minimum, 100, and range, [100, 1000], by the linear transformation

$$m^* = 100 + \frac{900[m - \min(m)]}{[\max(m) - \min(m)]} \quad (2)$$

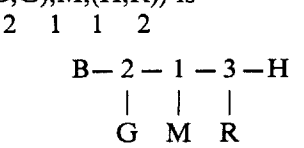
The overall quality and validity of the dissimilarity measures is assessed by a combination of three factors: (1) the F measure of goodness of fit of the best of all possible trees to the data, (2) whether the best tree is the same as the canonical tree, and (3) whether the nonnegative least squares fit of the tree has arbitrarily assigned a length of zero to some branches that otherwise would have been given unrealistic negative values by unconstrained least squares. The canonical tree is the best of all possible trees obtained for the conventional singlet mismatch count of aligned sequences. Interestingly, it is also the tree found for a plurality of all the other dissimilarity measures applied to all the various sequence codings. A determination of the relative weights of these three factors in the combination has not been attempted. All three are reported in the evaluation of all the similarity measures in Table 4 and of all the possible five-species trees in Table 1.

Results and Discussion

Table 1. Tables 1, 2, and 3 assess the suitability of the additive least squares method for inferring trees from dissimilarity measures, and the suitability of the F-value measure of the goodness of the fit obtained for assessing the quality of the dissimilarity measures. Table 1 shows that a ranking of goodnesses of fit to possible trees provides a sensitive measure of the validity of different dissimilarity measures (distances) derived from different sequences for the same set of species. It presents F-val-

ues and lists of branches of inferred length zero for each of the 15 possible trees for five mammalian species (human, rabbit, mouse, bovine, goat), using several different similarity measures: five for nonaligned noncoding sequences and two for aligned coding sequences, including the conventional singlet mismatch counts. There are several reasons for choosing these five mammalian species for the test. Increase of the number of species to six produces 105 possible trees, too many to justify complete tabulation. Data from another mammal were not available, and inclusion of an evolutionarily distant bird (chicken or duck) or amphibian (*Xenopus*) would produce a very large distance to all the mammalian species that would add little to a test of discriminability of the method among trees for the five more closely related mammalian species. For example, in Table 3 α-globin, the distance of *Xenopus* to the mammalian subtree, edge(X–2) + edge(2–4) = 609 + 184 = 793, and of chicken or duck, edge(C–3) or (D–3) + edge(3–2) + edge(2–4) = 47 or 53 + 131 + 184 = 362 or 368, is to be compared with the longest distance within the mammalian subtree, edge(M–4) + edge(4–5) = 178 + 44 = 222. Substitution of the distant species makes the fit to the best tree much better than to any of the 14 alternatives, and this reduces the sensitivity of discrimination among the dissimilarity measures.

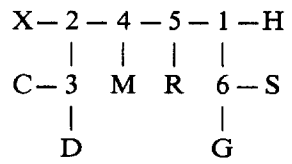
The trees are unrooted and are specified in the standard nested parentheses form (Felsenstein 1987). The leaves (tips, species) are designated by the one-letter codes given above. The internal nodes, all ternary and all equivalent in an unrooted tree, are designated by commas, here numbered for ready reference. The subtrees (including leaves) directly connected to a given node occur in pairs, separated by that node (comma) within a given level of parenthesization, except for one arbitrary internal node, designated node 1 in the tables, which occurs twice, separating three subtrees at the outermost level of parentheses. In Table 1, I have designated fitted branches of length zero by enclosing the nodes (including leaves) that they connect in single parentheses separated by dashes. Perhaps the more familiar graphical representation of the best tree in Table 1 ((B,G),M,(H,R)) is



and the α-globin tree

$$((((X_2,(C_3,D))_4,M)_5,R)_1,(S_6,G)_1,H)$$

in Table 3 is



The conventional count of singlet mismatches of aligned coding sequences gives high values of F and no enforced zero branches for trees 1 and 3, the F-value for 3 being somewhat higher, and it also gives a high value of F for tree 2 but with an enforced zero branch (1–3). Parenthetically, tree 2 is the best tree for α-globins using the conventional count of single mismatches (cf. Table 4). Those three values of F are 30–40 times greater than for any other tree, all of which have enforced zero branches. These three trees are the only ones in which B and G are connected through a single node, which therefore is determined to be a definitely established subtree. The connectivity (specification of all pairs of nodes directly connected by a single edge) of the (M,R,H) subtree is left ambiguous.

The remaining six dissimilarity measures find only one tree with an outstandingly high F-value and no enforced zero branch. For the triplet distance measures (MDD), one for coding and one for noncoding sequences, this high value is 130–800 times greater than for any other tree, all of which have enforced zero branches. For both measures, this high value is for tree 3, the tree having the somewhat higher of the two high values for the conventional dissimilarity measure. For the four longest common word measures (CLW), all for unaligned noncoding sequences, this high value is 4–70 times greater than for any other tree, all of which have enforced zero branches. For all four CLW measures this high value is for tree 1, the tree having the somewhat lower of the two high values for the conventional similarity measure.

It is concluded that the additive least squares method of fitting trees to distance measures has good discriminating power among trees for these relatively closely related mammalian globin sequences and thus provides a good assessment of the merits of the different similarity measures. The similarity measures not requiring sequence alignment, including applications to unalignable noncoding sequences, are about as effective for these purposes as is the traditional count of single mismatches of aligned coding sequences. It is anomalous that the former measures are so decisive in finding a single best tree, whereas the traditional method finds two exceptionally good trees of approximately equally good fit.

Table 2. Table 2 shows that the two trees given outstandingly good overall fits by a dissimilarity measure in Table 1 also yield closely similar esti-

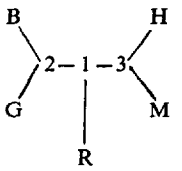
Table 1. Measures (F, zero edges)[a] of fit of several dissimilarity measures to all possible trees for β-globin sequences of five mammalian species[b]

Nodes and trees[c]	Dissimilarity measures[d]			
	Coding		Noncoding	
	Singlet mismatch	Triplet distance	Triplet distance	SIB (13)
2 1 1 3				
((B, G), R, (H, M))	340	26 (1–3)	9 (1–3)	9919
((B, G), H, (R, M))	326 (1–3)	26 (1–3)	9 (1–3)	137 (1–3)
((B, G), M, (H, R))	468	20,950	2883	137 (1–3)
((B, M), G, (H, R))	12 (1–2)	12 (1–2)	22 (1–2)	13 (1–2)
((G, M), B, (H, R))	12 (1–2)	12 (1–2)	22 (1–2)	13 (1–2)
((R, B), G, (H, M))	12 (1–2)	5 (1–2)	4 (1–2), (1–3)	20 (1–2)
((R, B), H, (G, M))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–2)	12 (1–2), (1–3)
((R, B), M, (H, G))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–2), (1–3)	12 (1–2), (1–3)
((B, M), R, (H, G))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–3)	12 (1–2), (1–3)
((R, M), B, (H, G))	10 (1–3)	5 (1–3)	4 (1–2), (1–3)	13 (1–3)
((R, G), B, (H, M))	11 (1–2)	5 (1–2)	4 (1–2), (1–3)	20 (1–2)
((R, G), H, (B, M))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–2), (1–3)	12 (1–2), (1–3)
((R, G), M, (H, B))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–2), (1–3)	12 (1–2), (1–3)
((G, M), R, (H, B))	10 (1–2), (1–3)	5 (1–2), (1–3)	4 (1–3)	12 (1–2), (1–3)
((R, M), G, (H, B))	10 (1–3)	5 (1–3)	4 (1–2), (1–3)	13 (1–3)

[a] Entries in the table are the value of the F-statistic of the least squares fit followed in parentheses by the pairs of nodes connected by an edge assigned value zero by the nonnegative least squares

[b] The raw similarity measures, m, have all been normed, m*, to the same range [100, 1000] by a linear transformation m* = 100 + [m − min(m)]·900/[max(m) − min(m)]. The measure of fit to the tree is the dimensionless ratio F = (variance of observed pair differences explained by nonnegative least squares fit to tree)/(variance not explained). Only the singlet mismatch similarity measure requires prior alignment of the sequences. Reasonable a priori alignment of the noncoding sequences is not attainable. The apparent equality of low F-values for different trees is often an artifact of rounding off. For example, for rows 14 and 15 of the singlet mismatch column, F-values are 9.76 and 10.20, respectively

[c] Trees are represented as follows: Terminal nodes (leaves) are designated by the critical letter of the species name. Internal nodes are designated by numbered commas. Two (or in one case three) nodes linked to the same node (comma) are enclosed in parentheses. For example, row 1 ((B, G), R, (H, M)) designates



[d] The dissimilarity measures are (1) singlet mismatches are the raw counts of the number of mismatched bases in a pair of aligned sequences, (2) triplet distances are the squared Euclidean distances between the 64-coordinate vectors of raw counts of the numbers of overlapping base triplets of two sequences, (3) SIB, SIS, and SAB distances are defined in the Methods section. SIB (13) is for the $P \leq 0.01$ significant common word length, 13. SIB (10) is for common words of length ≥ 10 where incompatibly aligned words of lengths 10, 11, or 12 have been removed on an ad hoc basis

mates of the several branch lengths (edges). It presents all estimated edges of trees for the five mammalian β-globins using the same set of seven similarity measures used in Table 1. Results are presented for the two trees found best by the conventional measure and by one or other of the new measures. The two triplet distance measures, one for coding and one for noncoding sequences, agree with each other and with the higher of the conventional measure F-values in their designation of the best tree. The F-values for the triplet distance measures are 6 and 45 times greater, for noncoding and coding, respectively, than for the conventional measure. The longest common word distances for noncoding sequences agree with each other and with the lower of the conventional measure F-values in their designation of the best tree. Their F-values are 2–29 times greater than for the conventional measure, increasing monotonically in the order SAB, SIS, SIB (10), and SIB (13), the last and highest being for exactly matching and uninterrupted significantly long words.

The vectors of estimated branch lengths are similar visibly for the seven dissimilarity measures. A more quantitative measure of their similarity is presented in the correlation matrix of Table 2b. The 5% significance value for the correlation coefficient of seven vectors is 0.669. The values of all pairs exceed this value except for pairs in the lower tree that include the noncoding triplet distance measure.

Table 1. Extended

Dissimilarity measures		
Noncoding		
SIB (10)	SIS	SAB
2687	1045	678
199 (1–3)	196 (1–3)	182 (1–3)
199 (1–3)	196 (1–3)	182 (1–3)
12 (1–2)	11 (1–2)	12 (1–2)
12 (1–2)	11 (1–2)	12 (1–2)
16 (1–2)	14 (1–2)	15 (1–2)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–3)	10 (1–3)	11 (1–3)
16 (1–2)	14 (1–2)	15 (1–2)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–2), (1–3)	10 (1–2), (1–3)	10 (1–2), (1–3)
11 (1–3)	10 (1–3)	11 (1–3)

However, the branch values for the triplet distance measures for coding and noncoding sequences correlate highly with each other and the former correlates highly with the values for the conventional measure. The highest correlations are among the four values of the longest common word measures (all $P < 0.0005$), the highest values being for the pair (SIS, SAB) having the lowest F-values among the four for the best tree.

It is concluded that the seven dissimilarity measures that are presented agree well in the branch lengths estimated from them for the best tree and for a less well-fitted tree. The measures include the conventional singlet mismatch count for aligned sequences and six measures not requiring alignment, one for alignable coding sequences and five for nonalignable noncoding segments.

Table 3. Table 3 shows that all the estimated edges in some selected fitted trees are determined to satisfactory significance levels. It contains the estimated branch lengths and the significance levels of them for the best trees derived from DNA sequence data for eight β- and eight α-globins using the singlet mismatch counts of aligned coding sequences and using the triplet distance measure for unaligned β coding sequences. The statistical significance of the individual estimated branch lengths can be determined by the additive least squares procedure applied to the incidence matrix corresponding to a given tree. For eight sequences there are $(2)(8) - 3 = 13$ edges and $= \binom{8}{2} = 28$ distances. The

Table 2a. Edges (branch lengths) estimated from normed dissimilarity measures of two selected trees for β-globin sequences of five mammalian species

	Estimated edges[a]						
	Coding		Noncoding				
	Singlet mismatch	Triplet distance	Triplet distance	Complement of SIB (13)	Complement of SIB (10)	Complement of SIS	Complement of SAB
Tree and nodes: ((B,G),R,(H,M)) 2 1 1 3							
Edge[b]	340	26	9	9919	2687	1045	678
B–2	15	57	54	43	45	46	45
G–2	85	43	46	57	55	54	55
H–3	191	449	485	418	464	430	435
M–3	626	402	327	547	536	530	543
R–1	227	309	263	330	313	294	322
1–2	315	437	364	289	332	342	340
1–3	12	0	0	110	91	83	84
Tree and nodes: ((B,G),M,(H,R)) 2 1 1 3							
Edge[b]	468	20,950	2883	137	199	196	182
B–2	15	57	54	43	45	46	45
G–2	85	43	46	57	55	54	55
H–3	182	340	323	467	505	467	473
R–3	214	200	102	342	323	303	331
M–1	628	375	287	596	576	568	581
1–2	308	355	243	325	362	370	368
1–3	33	245	363	0	0	0	0

[a] The pairwise dissimilarity measures have been normed as described in the legend to Table 1

[b] Edges are designated by a dash between the designations of the two nodes that they link. Edges have been estimated by nonlinear least squares

Table 2b. Correlations of vectors for seven dissimilarity measures of edges estimated for each of two selected trees for β-globins of five mammalian species

	Coding triplet distance	Noncoding triplet distance	Noncoding complement of SIB (13)	Noncoding complement of SIB (10)	Noncoding complement of SIS	Noncoding complement of SAB
Tree: ((B,G),R,(H,M))						
Coding singlet mismatch	0.743	0.633	0.875	0.849	0.876	0.876
Coding triplet distance		0.979	0.882	0.928	0.932	0.928
Noncoding triplet distance			0.848	0.906	0.900	0.894
Noncoding complement of SIB (13)				0.991	0.988	0.993
Noncoding complement of SIB (10)					0.998	0.998
Noncoding complement of SIS						0.999
Tree: ((B,G),M,(H,R))						
Coding singlet mismatch	0.705	0.329	0.864	0.835	0.858	0.858
Coding triplet distance		0.826	0.773	0.798	0.807	0.797
Noncoding triplet distance			0.379	0.409	0.407	0.393
Noncoding complement of SIB (13)				0.995	0.993	0.997
Noncoding complement of SIB (10)					0.998	0.998
Noncoding complement of SIS						0.999

incidence matrix has 28 rows corresponding to the observed distances and 13 columns corresponding to the 13 edges to be estimated. Each row has 1 in those columns whose sum theoretically equals the corresponding observed distance and 0 in the remaining columns. For example, in the α-globin tree diagrammed above, distance (H–R) = the sum of edges (H–1), (1–5), and (5–R). The equations are grossly overdetermined and suitable for least squares estimation. I used function "regress" in the S statistical package available in the UNIX operating system. The table presents the Student t-values from application of "regress" for all the branches for the best unrooted tree fitted to the distance data (dissimilarity measures). The best trees are the same for the two β-globin data sets using different similarity measures but are different for the α and β data sets using the same similarity measure. They differ in the subtrees for the five mammalian species. The β subtree corresponds to tree 3 of Table 1, the best tree designated by the singlet mismatch and triplet distance measures. The α subtree corresponds to tree 2 of Table 1, a tree found to be good by the singlet mismatch measure but very poor by all six other measures. All estimated branch lengths for both the α and β data for singlet mismatches and for the β data for both similarity measures are statistically significant ($P \leq 0.1$). The longest branch and most significant ($t \geq 20$) is that connecting the amphibian X to the tree. The shorter branches have lower significances ($t = 1.5$–4.3).

It is concluded that the triplet distance measure not requiring sequence alignment yields estimated lengths for all edges that are of satisfactory statistical significance and are as significant as those yielded by the conventional mismatch count measure requiring sequence alignment.

Table 4. Table 4a assesses the robustness of the results of application of the additive least squares method relative to choices from a variety of dissimilarity measures derived from a variety of data sets on the same set of sequences. Robustness is assessed by the congruence of three quality measures for the different choices of dissimilarity measures and data sets. The table presents the quality measures of fits of trees to aligned coding sequences of eight α- and eight β-globins and to the available unalignable noncoding sequences of six α- and seven β-globins.

All three measures of the quality of fit of the tree to the observed data are given for each sequence subset and each similarity measure. The three quality measures are the variance ratio measure of the goodness of fit, F; the listing of any edges arbitrarily set to zero length by the nonnegative least squares algorithm when fitting the canonical tree; and the specification of the best tree and its F-value in those cases where it is different from the canonical tree. Table 4b presents counts of cases in which the best tree is the same as the canonical tree (the best tree for aligned singlet mismatch counts) and of the cases in which it is different.

Results for raw base singlet mismatch counts on the aligned coding sequences are presented for 10 different subsets of the sequences, the whole sequences, codon sites 1, 2, 3, exons 1, 2, 3, and three kinds of substitutions: transitions, transversions preserving weak and strong hydrogen-bonding character, and transversions not preserving those characters. For the same whole sequences, results for raw mismatch counts are presented for four additional objects: base doublets, base triplets, codon singlets, and amino acid singlets. Fits for Jukes–Cantor (1969) corrected singlet mismatch counts are presented for the whole base sequences. The Jukes–

Table 3. Edges, estimated from normed dissimilarity measures, of the canonical trees for coding sequences of eight α- and eight β-globins[a]

α-globin (((X,(C,D)),M),R),(S,G),H)[b] 2 3 4 5 1 6 1			β-globin ((((X,(C,D)),M),(B,G)),R,H)[b] 2 3 4 5 6 1 1				
Singlet mismatch 1486[c]			Singlet mismatch 933[c]			Triplet distance 364[c]	
Edge	Length	t	Edge	Length	t	Length	t
X–2	609.2	62.7	X–2	621.5	48.5	518.3	21.9
C–3	46.6	3.7	C–3	44.5	2.7	52.5	1.7
D–3	53.4	4.3	D–3	55.5	3.4	169.5	5.5
2–3	131.2	9.0	2–3	236.5	12.3	294.4	8.3
M–4	178.2	20.1	M–4	225.8	19.3	118.0	5.5
2–4	184.1	15.5	2–4	127.2	8.1	118.6	4.1
R–5	92.4	10.4	G–6	57.2	3.5	55.1	1.8
4–5	44.1	3.9	B–6	42.8	2.6	44.9	1.5
G–6	44.3	3.5	4–5	34.5	2.5	92.2	3.6
S–6	55.7	4.4	5–6	96.8	5.5	128.7	4.0
5–1	88.5	9.1	H–1	101.1	6.1	119.9	3.9
1–H	32.3	2.7	R–1	106.3	6.4	178.4	5.8
1–6	23.4	1.6	1–5	27.2	1.5	62.0	1.9

[a] Legend as for Table 2a
[b] Tree and nodes
[c] F-value

Cantor formula corrects raw counts for multiple substitutions at the same base site and for parallel substitutions in the two sequences at the same base site on the assumption that all possible substitutions are equally likely and all base sites are equally substitutable.

For the same whole coding sequences, results are presented for 11 similarity measures not requiring sequence alignment: eight multiplet distribution distance measures (MDD) for base singlets, doublets, triplets, and quartets, amino acid singlets and doublets, and codon singlets and doublets, and three complements of significantly long common words measures (CLW) for SIBs, SISs, and SABs.

For the unalignable noncoding sequences, results are presented for the distribution distance measures (MDD) for base doublets, triplets, and quartets. No results for mismatch counts are presented because the noncoding sequences cannot be aligned satisfactorily. No results for complements of significantly long common words (CLW) are given because no significant common words exist between a bird or amphibian and a mammal for these highly divergent sequences. The canonical trees for the noncoding subsets of species are extracted from the canonical trees for the coding supersets of all eight species.

For the β-globin coding data, most of the similarity measures, both those requiring and those not requiring prior sequence alignment, find the best tree to be the canonical tree (20/26) (Table 4b). Four of the failures occurred with the conventional mismatch count method requiring sequence alignment and two with the new methods not requiring sequence alignment. In some cases where the best tree is not the canonical tree, the difference in F-values is small, and forcing a fit to the canonical tree does not produce branches of zero length (e.g., rows 13 and 14 for transitions and one class of transversions, respectively). In other cases the difference in F-values is large and branches of zero length are produced (e.g., row 15 for the other class of transversions). In all cases where the best tree is not the canonical tree, the canonical (X,(C,D)) subtree is retained. However, its insertion into the mammalian subtree is moved (rows 10 and 15). In the other cases the mammalian subtree itself is changed (rows 13, 14, 16, and 20 of Table 4 corresponding to mammalian trees 1, 15, 6, and 2 of Table 1, respectively).

For the seven available unalignable noncoding β-globin sequences the interpretation of the results is less clear. The canonical tree is assumed to be that obtained merely by removing the branch to the duck from the eight-species tree. Fitting this tree yields low F-values compared to the best tree and produces edges of zero length. The best trees for all three dissimilarity measures are the same (rows 27, 28, 29) and the connectivity of the mammalian species is the same as for the best tree (number 3 of Table 1). However, the integrity of the (X,C) subtree is not maintained; its members are inserted separately into the mammalian tree and neither into the same branch as in the best eight-species tree.

The results for the α-globin data are much less satisfying. Even for the coding data, the fraction of

Table 4a. Comparison of various dissimilarity measures for α- and β-globin sequences[a]

	Canonical tree		Best tree		Canonical tree		Best tree	
	F[a]	0-edges[a]	Code[a]	F[a]	F[a]	0-edges[a]	Code[a]	F[a]
	α-globin				β-globin			
	(((X,(C,D)),M),R),(G,S,H)[b] 2 3 4 5 1 6 1				(((X,(C,D)),M),(G,B)),R,H)[b] 2 3 4 5 6 1 1			
Coding regions: mismatch count measures needing prior alignment								
1, Base singlets	1486				993			
2, Base singlets + Jukes–Cantor	937				495			
3, Base doublets	1436				1075			
4, Base triplets	1119				1362			
5, Codon singlets	482				1185			
6, Amino acid singlets	576	(2–3)	((((X,(C,D)),R),M),(G,S),H)	639	485			
7, Base singlets codon site 1	445	(2–3)	((((X,(C,D)),R),M),(G,S),H)	479	846			
8, Base singlets codon site 2	412	( )	((((X,(C,D)),(M,R)),(G,S),H)	416	686			
9, Base singlets codon site 3	347	( )	(((((X,(C,D)),M),R),G),S,H)	357	730			
10, Base singlets in exon 1	283				508	(2–3)	((((X,(C,D)),G),B),M),R,H)	650
11, Base singlets in exon 2	574	(1–6)	((((X,(C,D)),M),R),S,(G,H))	583	646			
12, Base singlets in exon 3	523	( )	(((X,(C,D)),(R,(S,G))),M,H)	578	412			
13, Transition singlets T ↔ C, A ↔ G	328	(4–5)	((((C,(D,X)),M),R),(G,S),H)	344	230	( )	((((X,(C,D)),M),H),(G,B),R)	237
14, Transition singlets T ↔ A, C↔ G	374				559	( )	(((((X,(C,D)),M),R),G),B,H)	801
15, Transition singlets T ↔ G, C ↔ A	100	(1–2)	(((X,(C,D)),(G,M)),(R,S),H)	106	1396	(2–3)	((((X,(C,D)),(B,G)),M),R,H)	3132
Coding regions: measures not needing prior alignment								
16, Distribution of base singlets	118	(3–4) (4–5) (1–6)	(((((D,(X,M)),C),R),S),G,H)	615	174	(1–2)	((((X,(C,D)),M),((B,R),G),H)	312
17, Distribution of base doublets	119	(3–4) (4–5) (1–6)	(((((D,(X,M)),C),R),G),S,H)	528	272	(C–3)		
18, Distribution of base triplets	249	(4–5) (1–6)	(((((D,(X,M)),C),R),G),S,H)	482	364			
19, Distribution of base quartets	213				426			
20, Distribution of amino acid singlets	90	(2–3)	((((X,(C,D)),R),M),(G,S),H)	180	749	(1–2)	((((X,(C,D)),M),R),(B,G),H)	827
21, Distribution of amino acid doublets	265	(2–3)	((((X,(C,D)),R),M),(G,S),H)	305	3057			
22, Distribution of codon singlets	170	(1–2)	((((X,(C,D)),M),(S,G)),R,H)	171	312			
23, Distribution of codon doublets	327				1482			
24, Complements of SIBs	130	( )	((((X,(C,D)),M),(G,S)),R,H)	142	1882			
25, Complements of SISs	189	( )	((((X,(C,D)),M),(G,S)),R,H)	214	599			
26, Complements of SABs	104	( )	((((X,(C,D)),M),(G,S)),R,H)	112	379			
	α-globin (((C,D),M),(G,S),H)[b] 2 3 1 4 1				β-globin ((((X,C),M),(B,G)),R,H)[b] 2 3 4 5 1 1			
Noncoding regions: measures not needing prior alignment								
27, Distribution of base doublets	73	(1–4)	((((D,(C,M)),S),G,H)	77	58	(3–4)	(X,(((C,(B,G)),M),R),H)	886
28, Distribution of base triplets	62	(1–4)	((((C,D),S),M),G,H)	70	63	(2–3) (3–4)	(X,(((C,(B,G)),M),R),H)	971
29, Distribution of base quartets	47	(1–4)	((((C,D),M),S),G,H)	50	85	(2–3) (3–4)	(X,(((C,(B,G)),M),R),H)	1833
30, Complements of SIBs[c]								

Table 4b. Contingency table of counts of rows where the best tree equals the canonical tree

	α-globin				β-globin			
	Measures needing alignment		Measures not needing alignment		Measures needing alignment		Measures not needing alignment	
Sequence/count	=	≠	=	≠	=	≠	=	≠
Coding	7	8	2	9	11	4	9	2
Noncoding			0	3			0	3

cases in which the best tree is the same as the canonical tree is only about one-half for the dissimilarity measures requiring prior sequence alignment and much less for measures not requiring alignment (Table 4b). However, for the eight failed measures requiring prior sequence alignment, none of the differences in *F*-values is large, and in three cases no branches of zero length are produced. In all these cases the canonical (X,(C,D)) subtree is retained except transition singlet mismatches (row 13), where it is (C,(D,X)).

For the measures not requiring prior sequence alignment, base quartet distances and codon doublet distances find the best tree to be the canonical tree. Also for the amino acid singlet and doublet distances, the codon singlet distances, and the three complement of long word distances, the differences in *F*-values between the best and canonical tree are not large, the (X,(C,D)) subtree is retained, and for the complement of long word distances, no branches of zero length are introduced. In the 14 cases mentioned so far, many possible mammalian subtrees have arisen in the best trees, namely trees 1, 2, 3, 7, 10, and 15 of Table 1, and many changes have occurred in the insertion of the (X,(C,D)) subtree into them. The best trees for the base singlet, doublet, and triplet distance measures (rows 16, 17, 18) all give much larger *F*-values than do the canonical trees, and all have lost the separate (X,(C,D)) subtree. The mammalian subtree for row 16 is the same as that for base singlet mismatches in exon 2, row 11. The mammalian subtrees for rows 17 and 18 are the same as that for base singlet mismatches in codon site 3. Although the {X,C,D,M} subtree is different in rows 16, 17, and 18, its insertion into the {R,G,S,H} subtree is the same as in the canonical tree, namely into the branch to terminal R.

For the six available unalignable noncoding α-globin sequences, the best tree is different from the canonical tree in all three cases. The canonical tree is assumed to be that obtained merely by removing the branches to the rabbit and *Xenopus* from the eight-species tree. Fitting this tree produces edges of zero length. However, in no case is the difference in *F*-values large. The best trees differ for all three similarity measures, doublet, triplet, and quartet distances (rows 27, 28, 29). The triplet and quartet distance best trees both have retained the (C,D) subtree, but the mammalian subtrees are different.

Only one case of production of a branch of zero length occurred in the total of all 58 best tree fittings of all similarity measures to both α- and β-globins and for both alignable coding sequences and unalignable noncoding sequences, namely for base doublet distances for β-globin coding sequences (row 17). Application of the Jukes–Cantor correction for multiple and parallel substitutions to the singlet mismatch data of both α- and β-globin coding sequences yields the canonical tree but with substantially reduced *F*-values for both data sets. Saitou and Nei (1987) also have found that use of the Jukes–Cantor adjustment impairs the inference of trees. It appears that the assumption of equally likely substitutions of all kinds of bases at all base sites is so faulty as to make the correction an impediment rather than an aid in the inference of trees.

For α- and β-globin coding sequences and β-globin noncoding sequences, increasing the length of the multiplet whose pairwise distances are measured increases the *F*-values for the fits to the canonical tree and increases the likelihood of its being the best fit. This holds for increases of base singlets to quartets, amino acid singlets to doublets, and for codon singlets to doublets (rows 16–23). For β-globin coding sequences alone, similar increases of the lengths of the multiplets whose mismatches are counted in-

←

[a] Entries in the table are the *F*-values of the nonnegative least squares fit of the normed dissimilarity measures to the canonical tree followed by the designation of any edges assigned length zero. If the canonical tree does not give the highest value of *F*, there then follows the specification of the different tree that gives the highest value and the *F*-value of the fit to it

[b] Canonical tree and nodes

[c] There are no values for complements of SIBs because for α-globin, neither chick nor duck has a significantly long word in common with any of human, goat, horse, or mouse, although each of these sets has several in common within its own members, and because for β-globin, neither chick nor *Xenopus* has a significantly long word in common with any of the mammals, although the mammals have several in common among themselves. In fact, about ⅔ of the pair (B,G) is covered by significantly long common words

Table 5. Combination of base singlet mismatch data of α- and β-globins of the same seven species[a]

Data	Best tree	F	Best other tree	F
α	((((X,(C,D)),M),R),G,H)	1167	((((X,(C,D)),M),G),R,H)	927
β	((((X,(C,D)),M),G),R,H)	823	((((X,(C,D)),M),R),G,H)	654
$\alpha + \beta$	α tree	593	β tree	601

[a] The best other tree for the α-globin data is the best tree for the β-globin data and vice versa

creases the F-values for the fits (rows 1, 3, 4). For β-globin coding sequences alone, the progression SAB, SIS, SIB increases the F-values of the fits (Table 4, rows 24, 25, 26 and Table 1, top row). In fact, F for SIB is greater than that for any other measure on base sequences, including mismatch counts. These observations suggest using multiplets rather than singlets for making aligned mismatch counts and for calculating unaligned distribution distances.

It is concluded that different distance measures often do not yield the same best tree and even that the same distance measure, conventional or new, often does not yield the same best tree when applied to different codings of the same sequence or, using the same coding, to different reasonable selections of subsequences from it. However, in all these variations the new measures not requiring sequence alignment perform about as well as does the conventional measure requiring prior sequence alignment.

Table 5. It is obvious that there is only one historically correct evolutionary tree for a given set of species. In the event of the production of different trees from different sets of data, such as for α- and β-globins in the present work, or by the use of different methods of inference on the same set of data, such as additive least squares and parsimony, or by the use of different dissimilarity measures on the same set of data, how can these different trees be combined to estimate an overall best tree and how can its significance be evaluated? Program CONSENSE in the PHYLIP package of Felsenstein (1986) takes as input the set of different trees and produces as output a list of the number of times each frequent subtree occurs in the set and a complete tree incorporating the frequent subtrees. The method does not use any information about how well the different trees have fitted their respective data sets, nor does it provide alternative good consensus trees with measures of their relative significance. The additive least squares method can do all of these things. Merely solve the least squares equation $AX = Y$ where Y is the concatenation of the vectors of observed pairwise distances for the different data sets, A is the rowwise combination of the incidence matrices (described above for Table 3) for one of the trees in the set of noncongruent trees, and X is the vector of computed branch lengths for the connectivity of the chosen tree, using the distance data from all observed data sets. The results for the different trees for the singlet mismatch distances of the seven species common to the α- and β-globin sets are presented in Table 5. In this case there are $\binom{7}{2} = 21$ pairwise distances so that Y is a 42 vector. There are $2 \times 7 - 3 = 11$ branches so that X is an 11 vector and A is a 42×11 matrix. The best tree derived from the β incidence matrix alone fits the combined data better than does the best tree derived from the α incidence matrix alone, but only very little better ($601 > 593$).

I give a parenthetical endorsement of the multiplet distance similarity measure. A clustering of doublet distances for a set of 30 varied genes surprisingly lumped some insulin genes with globin genes (Blaisdell 1986). This cluster was supported by some other multiplet measures, counts of common 7 words and of common 13 words with three mismatches, but no biological interpretation was proffered. I have found out since that the insulin and globin genes are close neighbors on the short arm of human chromosome 11 (Kittur et al. 1985). This similarity of insulin and globin base sequences and their closeness on the chromosomes are supportive of the widely accepted hypothesis that many current genes are the mutated descendants of ancestral duplications.

Acknowledgments. This work was supported in part by NIH grant R01GM3056-02. I thank S. Karlin, Y. Rinott, and an anonymous reviewer for helpful comments and M. Askew for preparation of the manuscript.

References

Astolfi P, Kidd KK, Cavalli-Sforza LL (1981) A comparison of methods for reconstructing evolutionary trees. Syst Zool 30:156–169

Blaisdell BE (1983) A prevalent persistent global nonrandomness that distinguishes coding and noncoding eucaryotic nuclear DNA sequences. J Mol Evol 19:122–133

Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA 83:5155–5159

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 32:550–570

Cornish-Bowden A (1979) How reliably do amino acid composition comparisons predict sequence similarities between proteins? J Theor Biol 76:369–386

Dayhoff MO (1979) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington DC, p 8

Dickerson RE, Geis I (1983) Hemoglobin: structure, function, evolution and pathology. Benjamin/Cummings, Menlo Park CA, p 93

Felsenstein J (1982) Numerical methods for inferring evolutionary trees. Q Rev Biol 57:379–404

Felsenstein J (1986) PHYLIP—phylogeny inference package (version 2.9). University of Washington, Seattle

Felsenstein J (1987) PHYLIP Newsletter, number 9, May 1987

Gibbs AJ, Dale MB, Kinns HR, MacKenzie HG (1971) The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acid sequence. Syst Zool 20:417–425

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro NH (ed) Mammalian protein metabolism. Academic Press, New York, pp 22–123

Karlin S, Ghandour G, Foulser DE, Korn LJ (1984) Comparative analysis of human and bovine papillomaviruses. Mol Biol Evol 1:357–370

Karlin S, Morris M, Ghandour G, Leung M (1988a) Efficient algorithms for molecular sequence analysis. Proc Natl Acad Sci USA 85:841–845

Karlin S, Morris M, Ghandour G, Leung M (1988b) Algorithms for identifying local molecular sequence features. CABIOS 4: 41–51

Karlin S, Ost F, Blaisdell BE (1989) Patterns in DNA and amino acid sequences and their statistical significance. In: Waterman MS (ed) Mathematical methods for DNA sequences. CRC Press, Boca Raton FL (in press)

Kittur SD, Hoppener JWM, Antonarakis SE, Daniels JDJ, Meyers DA, Maestri NE, Maarten J, Korneluk RG, Nelkin BD, Kazazian HH (1985) Linkage map of the shortarm of chromosome 11: location of the genes for catalase, calcitonin and insulin-like growth factor II. Proc Natl Acad Sci USA 82: 5064–5067

Konkel DA, Maizel JV, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosome beta-globin genes. Cell 18:865–873

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:408–425

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Sourdis J, Nei M (1988) Relative efficiencies of the maximum parsimony and distance-matrix methods in attaining the correct phylogenetic tree. Mol Biol Evol 5:298–311

Received November 8, 1988/Revised and accepted June 4, 1989