J Mol Evol (1987) 25:351–360

Journal of
Molecular Evolution

Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees

Da-Fei Feng and Russell F. Doolittle

Department of Chemistry, University of California–San Diego, La Jolla, California 92093, USA

Summary. A progressive alignment method is described that utilizes the Needleman and Wunsch pairwise alignment algorithm iteratively to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The sequences are assumed a priori to share a common ancestor, and the trees are constructed from difference matrices derived directly from the multiple alignment. The thrust of the method involves putting more trust in the comparison of recently diverged sequences than in those evolved in the distant past. In particular, this rule is followed: "once a gap, always a gap." The method has been applied to three sets of protein sequences: 7 superoxide dismutases, 11 globins, and 9 tyrosine kinase-like sequences. Multiple alignments and phylogenetic trees for these sets of sequences were determined and compared with trees derived by conventional pairwise treatments. In several instances, the progressive method led to trees that appeared to be more in line with biological expectations than were trees obtained by more commonly used methods.

Key words: Multiple sequence alignments — Evolutionary trees

Introduction

The evolutionary relationships of sets of protein (or nucleic acid) sequences are commonly depicted in the form of trees (Fitch and Margoliash 1967; Dayhoff et al. 1972; Moore et al. 1973; Sankoff et al. 1982; inter alia). Indeed, the digital nature of sequence data makes them more amenable to such treatment than is the case with many more qualitative biological characters. Most current schemes for constructing trees from sequences use a simple difference matrix, the elements of which are assembled by performing pairwise comparisons of all the sequences under study (Fitch and Margoliash 1967). A topology is found by classifying the sequences according to their differences, which ought to be a reflection of the evolutionary distances among them. For the most part, the principle of parsimony is rigorously adhered to, and the best trees are thought to be those that can account for the extant sequences by the smallest number of genetic events. The two important features of a tree are its topology, or branching order, and its branch lengths, which ought to be proportional to the true evolutionary distances.

In principle, the construction of an evolutionary tree based on sequence data ought to be a simple matter: all one has to do is cluster the sequences according to their similarities. In practice, uncertainties and ambiguities concerning both the topology and branch lengths are common, and enormous effort is often expended in finding the "best tree" (e.g., Fitch 1977; Penny and Hendy 1986). Finding the correct tree should depend on assembling a matrix that best describes the differences among the sequences, and this depends, in turn, on properly aligning the sequences (Hogeweg and Hesper 1984). The alignments can be obtained either by schemes that maximize similarity (Needleman and Wunsch 1970) or with those that minimize differences (Sellers 1974). If a similarity scheme is used, the scores must be transformed appropriately into measures of distance.

Offprint requests to: Da-Fei Feng

Ordinarily, alignments of either type are performed pairwise. The problem is that when the various paired alignments are grouped, they are seldom consistent one to another. Thus, when sequence A is paired with sequence B, gaps may appear at various locations, but when either A or B is aligned with a third sequence, C, the arrangement of gaps may be entirely different. Heretofore, this problem has been circumvented by making a multiple alignment of all the sequences by the judicious shifting of the sequences as needed to minimize differences ("eyeball" alignment).

The flaw in the approach is that these multiple alignments have, like pairwise alignment schemes before them, been subject to rigorous attempts at parsimony. Obviously, the closer two sequences resemble each other, the more confidence one has in the alignment. But in most multiple alignment schemes where maximum parsimony is sought, no distinction is made with regard to the confidence one has in a particular pairwise alignment. It seems to us folly that a gap should be discarded in an alignment of two closely related sequences merely because an alignment with some distantly related sequence might be improved.

To this end, we have devised a scheme of progressive sequence alignment that has a higher intrinsic regard for recent events than for distant ones. It is still based on a maximization of similarities, but it follows the simple rule "once a gap, always a gap." It is able to accomplish this by inserting neutral elements into sequences once gaps have been established. The sequences are aligned progressively, beginning with the most similar pair and continuing with the addition of the next most similar sequence or set of sequences. The difference scores obtained from the final alignment of all sequences are then used to construct the evolutionary tree. Ambiguities may still arise, of course, since the preliminary matrix of similarities (or differences) based on pairwise comparisons will often include what we call "better but less reliable" scores. These can be sorted out by testing alternative trees. Because it is impractical to consider all possible pairwise orders, we have adopted an effective compromise whereby reasonable alternative arrangements are explored progressively.

In this paper we describe the details of the method and apply it to several groups of protein sequences. Trees constructed by this approach can differ significantly from those assembled by traditional schemes, but they are often in accord with what might be expected on the basis of organismic phylogenies. The method has the added virtue of providing multiple sequence alignments quickly and simply by completely objective criteria.

Methods

Studies were performed on a DEC 11/730 VAX computer with the UNIX (Berkeley 43) operating system. The plotting package for use with a Nicolet Zeta plotter was written by Steve Dempsey of the U.C.S.D. Chemistry Department Computer Center. All utility programs were written in the C programming language (Kernighan and Ritchie 1978). The ensemble of programs dealing with sequence alignment and tree building can be contained by sending a blank magnetic tape to the authors.

Definitions. For purposes of description only, we would like to distinguish between *simple* and *compound* trees. Simple trees are those in which the branching order follows the simple clustering (((AB)C)D) etc., whereas compound trees have subclusters, as in ((AB)(CD)E). *Neutral elements* are simply characters (Xs) that are filled into sequences when gaps occur. They are neutral in the sense that they are invisible to the scoring system used to establish subsequent alignments, which is to say when X is matched with any other residue, the value is equal to zero. *Negative segments* are those internodal connecting distances with negative values that occasionally emerge from Fitch-Margoliash trees when data scatter confounds the segment averaging (or least-squares treatment). *Percent identity* is taken as the number of identities per 100 aligned residues.

Sequences. Amino acid sequences were taken from an updated version of the NEWAT database (Doolittle 1981). Primary references to the nine tyrosine kinase sequences and nine of the globin sequences have been provided in an earlier study (Feng et al. 1985). The additional globins used in the present study are from lamprey (Zelenik et al. 1979) and the bacterium *Vitreoscilla* (Wakabayashi et al. 1986). The superoxide dismutase sequences studied are human (Jabusch et al. 1980), bovine (Steinman et al. 1974), swordfish (Rocha et al. 1984), fruitfly (Lee et al. 1985), maize (Cannon et al. 1987), yeast (Johansen et al. 1979), and photobacter (Steffens et al. 1983).

Pairwise Alignments. The algorithm of Needleman and Wunsch (1970) was used in a three-matrix form (Fredman 1984) and utilized the Mutation Matrix of Dayhoff et al. (1978) in its scoring. The algorithm was actually employed in several slightly different settings. In the first, a program called SCORE aligns pairs of sequences in the conventional way and stores their alignment scores in a table. The similarity scores obtained from the alignments are converted to difference scores by the relationship

$$D = -\ln S_{eff} \times 100 = -\ln \frac{S_{real} - S_{rand}}{S_{ident} - S_{rand}} \times 100$$

where S_{real} is the alignment score itself, S_{rand} is the score obtained with random sequences of the same lengths and compositions, and S_{ident} is the average score of the two sequences being compared when each is aligned with itself. In practice, in these initial pairwise comparisons we use an average value for S_{rand} based on many previous observations (Feng et al. 1985). Inasmuch as this initial set of comparisons is assumed to be imperfect, no precision is lost by the modification, and considerable time is saved by the omission of numerous jumble comparisons. The value used, after normalization to a standard length, was 770, the average random score for numerous comparisons of many different kinds of sequences (Feng et al. 1985).

The Needleman-Wunsch algorithm is used in a second series of alignments in a mode in which gaps are concurrently filled with neutral elements. In the main version, DFalign, sequences are aligned successively. Should the tree in question be a com-

pound tree, subclusters are first prealigned with a simpler version of the program called PREalign.

Tree Building. A program based directly on the Fitch and Margoliash (1967) procedure was written in our laboratory by Mark Johnson. The program, BORD, was used to establish preliminary branching orders. Simply put, the smallest difference score is identified and a new matrix constructed that contains the average distances between members of the first pair and remaining members of the set. The procedure is repeated until all scores have been incorporated. A second program, BLEN, was used for determining branch lengths of the final tree. This program employs a least-squares approach as described by Klotz and Blanken (1981). In the event that a tree contains one or more "negative segments," the "nearest alternative" trees are considered and their scores compared. *Nearest alternative* trees are those in which the branches immediately adjacent to a negative segment are switched. The program TREEplot, also written by Mark Johnson, puts the data in an appropriate form for the Zeta plotter in order that dendrograms can be issued directly.

Outline of the Progressive Method

Pairwise Alignments

For n sequences, the number of pairwise alignments required for the initial matrix amounts to $(n - 1) \times n/2$. To this end, a simple UNIX shell program was constructed for running each comparison serially with the program SCORE; the resulting difference scores are automatically stored in a suitable file.

Identification of Most Closely Related Pair

The program BORD takes the output from SCORE and establishes a preliminary order of the sequences. The program BLEN uses the difference matrix from the SCORE program combined with a simple "connectivity table" to give branch lengths; the connectivity table merely puts all the connecting segments in tabular form. BLEN is only used at this point if trees based on pairwise comparisons are going to be prepared. The BORD program reveals whether or not the starting tree is simple or compound. In the case of compound trees, subclusters are prealigned with the program PREalign, which aligns the cluster and fills the gaps with neutral elements (Xs).

Progressive Insertion of Neutral Elements

The program DFalign, which is the heart of the procedure, is used to generate the multiple alignment. It begins by inserting neutral elements (Xs) in any gaps that occur in the aligned pair with the highest similarity score. After the original pair has been established and the gaps fixed, the next nearest relative or set of relatives is brought in and a new alignment made and a score determined. The key to this alignment is that new gaps can be incorporated into either sequence, but the earlier gaps are preserved. The first ternary arrangement, ABC, is then compared with the alternative BAC, the higher score being used to set the path for the next alignment. Similarly, when the next sequence is brought in, the arrangement ABCD is scored and compared with ABDC. Prealigned subclusters are maintained as separate units, however. The procedure is continued until all sequences have been incorporated.

Scoring the Final Alignment

The final alignment is scored with a modified regimen that recognizes the fixed nature of the gaps. Moreover, because the gaps are fixed, it is unnecessary to use an alignment program at this stage. Instead, a scoring system is used that measures S_{real} and S_{ident} in the usual way, but that employs a program, SHUFFLE, for determining S_{rand}. SHUFFLE randomizes each sequence numerous times while holding the gaps constant.

Constructing the Tree

The program BORD is used to obtain the new branching order and the program BLEN to determine the branch lengths. If any negative segments result, alternative trees with the branches on either side of the negative segment reversed are constructed and a new set of branch lengths calculated. If negative segments are still present, the alternating procedure is continued until they disappear, although we have not yet encountered a situation where more than one switch was necessary. The program TREEplot is used to produce the final dendrogram. A schematic outine of the programs called from start to finish is present in Fig. 1.

Results

Superoxide Dismutase

The sequences of seven copper–zinc superoxide dismutases—human, bovine, swordfish, fruitfly, maize, yeast, and photobacter—were subjected to a conventional pairwise alignment scheme and a tree constructed by the Fitch and Margoliash (1967) procedure (Fig. 2a). The same seven sequences were then treated by the progressive procedure and a tree generated (Fig. 2b). The trees differ both in branch order and branch length. More to the point, the progressive procedure yields a tree that corresponds to the accepted phylogeny of the organisms, whereas the conventionally generated tree does not.

In fact, the initial tree issued from the ordinary Fitch and Margoliash (1967) treatment had the expected phylogenetic branching order, but contained

a negative segment. When the nearest alternative tree was examined, generated by reversing the branches on either side of the negative segment, the sum of branch lengths was lowered, and a "better tree" with no negative segments emerged (Fig. 2a). The tree contradicts what is known of the evolutionary relationships of the organisms involved, however, in that the branch to the yeast sequence comes off above the branch to the *Drosophila* sequence.

Progressive Alignment Procedure

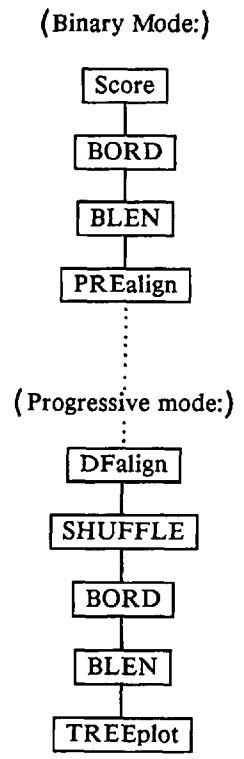

Fig. 1. Flow chart of progressive alignment procedure. Program names are shown in boxes. The program BLEN in upper portion of figure may be omitted if a tree based on pairwise alignments is not going to be constructed.

It should be emphasized that the multiple alignment (Fig. 3) used to obtain the final tree was obtained by strictly objective criteria and without recourse to "eyeball" manipulation. Moreover, the overall similarities, as reflected in the percent identities, are more in line with the true distances separating the organisms than are those observed in the original pairwise alignments (Table 1).

Hemoglobins

Eleven different globin sequences covering a broad spectrum of types were subjected to pairwise alignments and an initial tree constructed from the resulting difference matrix (Fig. 4a). The tree was similar to those presented in previous reports in that cyclostome globins (hagfish and lamprey) branch off in advance of the myoglobin–hemoglobin α-chain divergence (Goodman et al. 1974; Hunt et al. 1978; Feng et al. 1985). When the same 11 sequences were subjected to the progressive alignment procedure, the tree that emerged reversed the order to the more biologically reasonable situation in which the cyclostome globins are clustered with those of other vertebrates (Fig. 4b).

Also of interest are the relative positions of the plant and invertebrate hemoglobins. In the tree obtained from pairwise alignments, the plant and bacterial hemoglobins appear to be more closely related

Table 1. Percent identities calculated from binary (upper triangle) and progressive (lower triangle) alignment methods

	Superoxide dismutases						
	sodh	sodb	sodl	sddm	sdmz	sods	sdpb
sodh		82	67	60	62	53	31
sodb	82		74	57	61	55	35
sodl	67	72		59	59	56	35
sddm	59	59	58		68	54	31
sdmz	60	60	58	68		57	32
sods	51	52	54	51	54		30
sdpb	31	35	34	31	34	31	

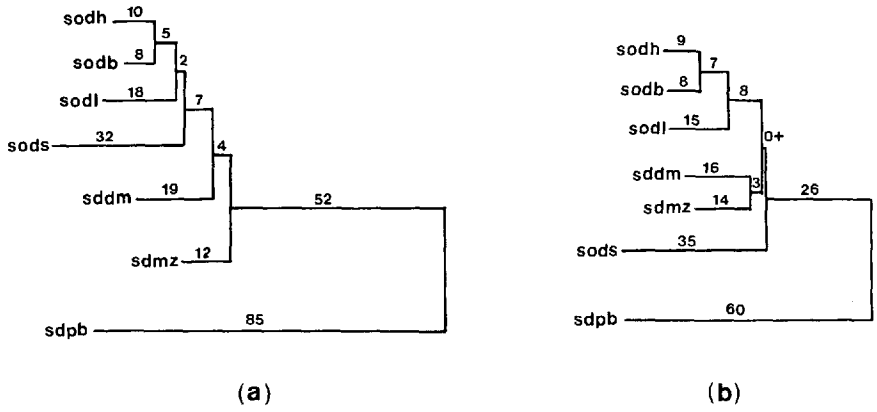

Fig. 2. Phylogenetic trees for seven superoxide dismutases, as determined by **a** simple pairwise alignments and **b** progressive multiple alignment. The four-letter designations are sodh, human; sodb, bovine; sodl, swordfish; sddm, fruitfly; sods, yeast; sdmz, maize; sdpb, photobacter. The same designations are used in Fig. 3 and Table 1.

```
                *     *                                ***  *   *        *          * *  *
sodh       ATKAVCVLKGDGPVQGSINFEQKESDGPVKVWGSIKGLTE GLHGFHVHQFG    NDTAGCT    SAGPHFNP LSRK
sodb       ATKAVCVLKGDGPVQGTIHFEAK  GDTVVVTGSITGLTE GDHGFHVHQFG    DNTQGCT    SAGPHFNP LSKK
sodl       VLKAVCVLRGAGETTGTVYFEQEGNANAVGKGIILKGLTP GEHGFHVHGFG    DNTNGCI    SAGPHFNP ASKK
sddm       VVKAVCVING  DAKGTVFFEQESSGTPVKVSGEVCGLAK GLHGFHVHEFG    DNTNGCM    SSGPHFNP YGKE
sdmz       MVKAVAVLAGT DVKGTIFFSQEGDG PTTVTGSISGLKP GLHGFHVHALG    DTTNGCM    STGPHFNP VGKE
sods     VQAVAVLKGDAG   VSGVVKFEQASESEPTTVSYEIAGNSPNAERGFHIHEFG    DATNGCV    SAGPHFNP FKKT
sdpb    QDLTVKMTDLQTG  KPVGTIELSQNKYG  VVFTPELADLTP GMHGFHIHQNGSCASSEKDGKVVLGGAAGGHYDPEHTNK

        *  *     * **                         *       *     *   *              *  * *  ** *
sodh    HGGPKDEERHVGDLGNVTADKDGVADVVIEDSVISLSGDHCIIGRTLVVHEKADDLGKGGNEESTKTGNAGSRLAcGVIGIAQ
sodb    HGGPKDEERHVGDLGNVTADKNGVAIVDIVDPLISLSGEYSIIGRTMVVHEKPDDLGRGGNEESTKTGNAGSRLAcGVIGIAK
sodl    HAGPKDEDRHVGDLGNVTADANGVAKIDITDK ISLTGPYSIIGRTMVIHEKADDLGRGGNEESLKTGNAGSRLAcGVIGTE
sddm    HGAPVDENRHLGDLGNIEATGDCPTKVNITDSKITLFGADSIIGRTVVVHADADDLGQGGHELSKSTGNAGARIGcGVIGIAK
sdmz    HGAPEDEDRHAGDLGNVTAGEDGVVNVNITDSQIPLAGPHSIIGRAVVVHADPDDLGKGGHELSKSTGNAGGRVAcGIIGLQG
sods    HGAPTDEVRHVGDMGNVKTDENGVAKGSFKDSLIKLIGPTSVVGRSVVIHAGQDDLGKGDTEESLKTGNAGPRPAcGVIGLTN
sdpb    HGFPWTDDNHKGDLPALFVSANGLATNPVLAPRLTL   KELKGHAIMIHAGGDNHS    DMPKALGGGGARVAcGVIQ
```

Fig. 3. Multiple alignment of seven superoxide dismutases determined by progressive method. Asterisks denote locations where all seven residues are identical. See legend to Fig. 2 for four-letter designations.

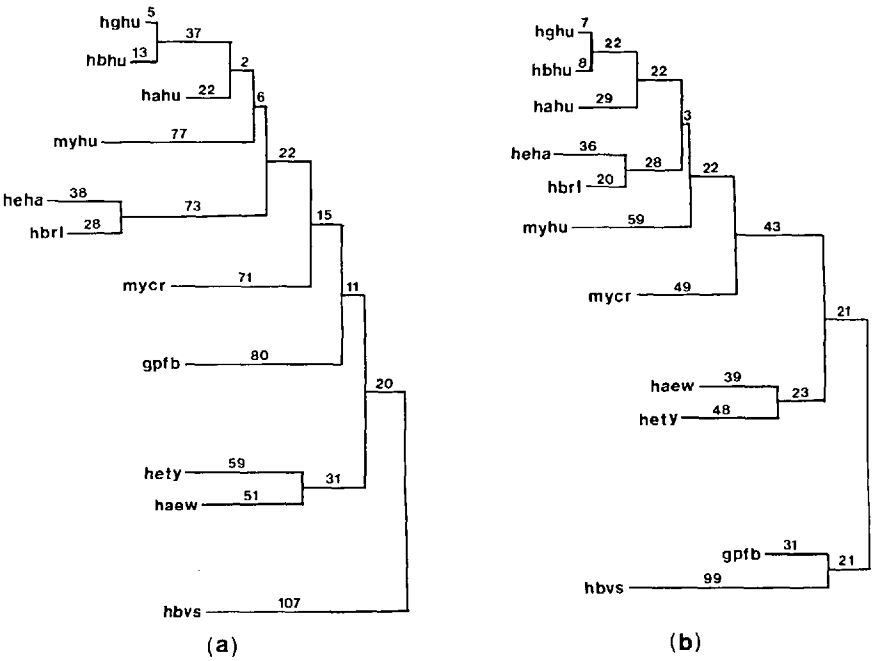

Fig. 4. Phylogenetic trees for 11 globin sequences as determined by a simple pairwise alignments and **b** the multiple alignments shown in Fig. 5. The four-letter designations are hghu, human globin γ chain; hbhu, human globin β chain; hahu, human globin α chain; heha, hagfish hemoglobin; hbrl, lamprey hemoglobin; myhu, human myoglobin; mycr, gastropod myoglobin; hety, earthworm hemoglobin (*Tylorrhynchus*); haew, earthworm hemoglobin (*Lumbricus*); gpfb, kidney bean leghemoglobin; hbvs, bacterial hemoglobin (*Vitreoscilla*). The same designations are used in Fig. 5 and Table 2.

to the globins of higher invertebrates and vertebrates than are those from annelid worms (Fig. 4a). Again, a more traditional grouping is obtained with the progressive alignment procedure (Fig. 4b). The multiple alignment generated by the procedure (Fig. 5) appears to be an accurate depiction of the history of events during globin evolution, and the degrees of similarity of the various globins based on these alignments are also more in line with expectations than are those found from simple binary alignments (Table 2).

Tyrosine Kinase-like Sequences

We had previously aligned a set of nine tyrosine kinase-like sequences and constructed a tree based on a simple pairwise matrix (Feng et al. 1985), and it was naturally of interest to see how the progressive alignment treatment compared (Table 3). In this case, unlike the situations with the superoxide dismutases and hemoglobins, the branching orders found by the two procedures did not differ (Fig. 6a and b). The multiple sequence alignment that was generated automatically during the procedure (Fig. 7) was somewhat different from the "eyeball" alignment made previously on the basis of a series of pairwise comparisons, although the same 14 invariant residues occur coincidentally in both renditions (Fig. 7). The trees themselves are not significantly different, although the branch lengths differ slightly.

Discussion

The concept of using pairwise alignments iteratively to establish phylogenetic relationships is hardly new. Moore et al. (1973) constructed the best possible dendrogram for a set of sequences by an iterative

```
                                                        *     *
hghu                GHFTEEDKATI  TSLW   GKV  NVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK VKAHGKKVLTSLG
hbhu                VHLTPEEKSAV  TALW   GKV  NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK VKAHGKKVLGAFS
hahu                 VLSPADKTNV  KAAW   GKVGAHAGEYGAEALERMFLSFPTTKTYFPHF DLSH     GSAQ VKGHGKKVADALT
heha        PITDHGQPPTLSEGDKKAI  RESW   PQIYKNFEQNSLAVLLEFLKKFPKAQDSFPKFSAKKS  HLEQDPA VKLQAEVIINAVN
hbrl        PIVDSGSVAPLSAAEKTKI  RSAW   APVYSNYETSGVDILVKFFTSTPAAQEFFPKFKGMTSADQLKKSAD VRWHAERIINAVN
myhu                GLSDGEWQLV  LNVW   GKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED LKKHGATVLTALG
mycr                SLQPASKSAL  ASSWKTLAKDAATIQNNGATLFSLLFKQFPDTRNYFTHFGNM SDAEMKTTGV GKAHSMAVFAGIG
haew                KKQCGVLEGLKVKSEWGRAYGSGHDREAFSQAIWRATFAQVPESRSLFKR     VHGDHTSDPA FIAHAERVLGGLD
hety                 TDCGILQRILVLQQWAQVYSVGESRTDFAIDVFNNFFRTNPD RSLFNR     VNGDNVYSPE FKAHMVRVFAGFD
gpfb                 GAFTEKQEALVNSSW EAFK  GNIPQYSVVFYTSILEKAPAAKNLFSF    LANGVDPTNPK LTAHAESLFGLVR
hbvs                  MLDQQTINIIKATV PVLK  EHGVTITTTFYKNLFAKHPEVRPLFD     MGRQESLEQPKALAMTVLAAAQNIE

hghu        DAIKHLD   DLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMV  TGVASALSSRYH
hbhu        DGLAHLD   NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVV  AGVANALAHKYH
hahu        NAVAHVD   DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFL  ASVSTVLTSKYR
heha        HTIGLMDKEAAMKKYLKDLSTKHSTEFQVNPDMFKELSAVFVSTM         GGKAAYEKLF  SIIATLLRSTYDA
hbrl        DAVASMDDTEKMSMKLRDLSGKHAKSFQVDPQYFKVLAAVIADTV        AAGDAGFEKLM  SMICILLRSAY
myhu        GILKKKGHHE   AEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKAL  ELFRKDMASNYKE LGFQG
mycr        SMIDSMDDADCMNGLALKLSRNHIQRKIGASRFGE MRQVFPNFLDEALGGGASGDVKGAWDALL  AYLQDNKQA  QA L
haew        IAISTLDQPATLKEELDHLQVQHEGRKIPDNYFDA FKTAILHVVAAQLGERCYSNNEEIHDAIACDGFARVLPQVLERG IKGHH
hety        ILISVLDDKPVLDQALAHYAAFH  LQFGTIPFKA FGQTMFQTIAEHI      HGADIGAWRAC  YA   EQIVT G ITA
gpfb        DSAAQLRANGAVVAD AALGSIHSQKGVSNDQFLV VKEALLKTLKQAV      GDKWTDQLSTALELA  YDELAAAI KKAYA
hbvs        NLPAILPAVKKIAVKHCQAGVAAAHYPIVGQELLGAIKEVLGDAATDDI      LDAWGKAYGVIADV   FIQVEADLYAQAVE
```

Fig. 5. Multiple alignment of 11 globins determined by progressive method. Asterisks denote locations where all 11 residues are identical. The order of the sequences was strictly based on the permutative trial described in the text. See legend to Fig. 4 for four-letter designations.

Table 2. Percent identities calculated from binary (upper triangle) and progressive (lower triangle) alignment methods

	Globins										
	hghu	hbhu	hahu	heha	hbrl	myhu	mycr	haew	hety	gpfb	hbvs
hghu		73	42	29	28	24	22	17	16	25	17
hbhu	73		45	26	24	25	22	18	18	23	20
hahu	42	45		20	35	27	24	22	19	15	18
heha	27	25	25		44	20	22	21	17	17	18
hbrl	26	24	34	44		23	19	26	23	15	14
myhu	25	25	28	18	23		22	21	17	17	19
mycr	21	21	23	19	18	21		18	20	15	25
haew	17	14	15	15	15	12	18		34	20	16
hety	16	15	14	12	14	12	18	34		16	13
gpfb	17	19	14	18	16	14	15	17	15		24
hbvs	11	11	12	6	10	10	10	11	12	15	

pairwise process, and, more recently, Hogeweg and Hesper (1984) used a heuristic approach for generating trees that also depends on successive pairwise alignments. As far as we know, however, the notion of "once a gap, always a gap," coupled with progressive pairwise alignment, has not been utilized before. Gap preservation is achieved by the insertion of neutral elements that hold the gap positions fixed during each progressive realignment.

Two things are certain: the method, while heuristic, provides multiple sequence alignments that are based on objective criteria, and trees derived from these alignments appear to be in harmony with the biology of the proteins as evidenced by the phylogeny of the organisms from which they are obtained. The simplicity of the procedure is attested to by the small number of pairwise comparisons that must be undertaken to produce the multiple alignment (Table 4). Thus, if 10 sequences are to be aligned, only 61 comparisons have to be made. This is a smaller number of alignments than is ordinarily performed when a set of jumbles is made for a single quantitative alignment. In this regard, we have eschewed the use of jumbled comparisons in the initial alignments in favor of an empirically determined average random score.

Kinds of Sequence Alignment

Broadly speaking, there are three kinds of multiple sequence alignment: (1) structural equivalence types, (2) global optimization methods, and (3) historical alignments. The first of these, structural equivalence, is used mainly by crystallographers. The goal is to align those segments of two protein sequences that occupy equivalent three-dimensional orientations. As such, these studies are usually restricted to protein families at least one member of which has had an x-ray structure determined (Bajaj and Blundell 1984). The interest is focused on present-

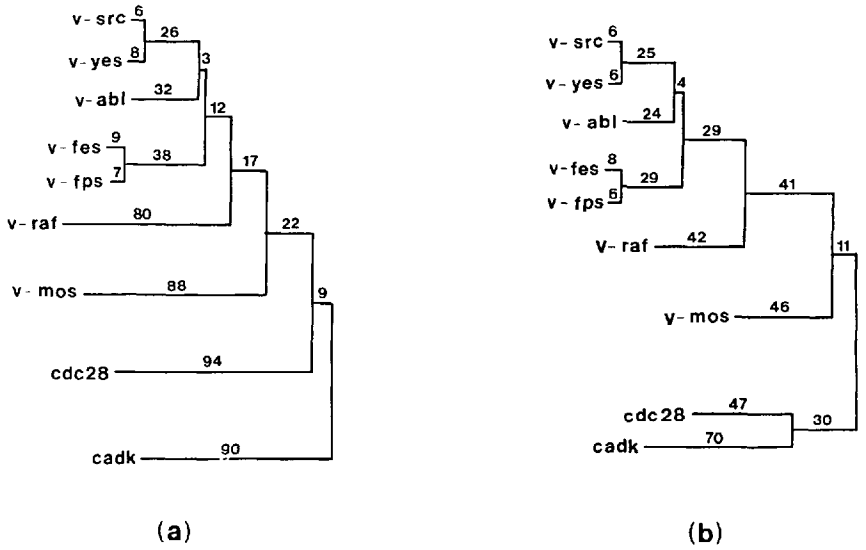

Fig. 6. Phylogenetic trees for nine tyrosine kinase-like sequences determined from a simple pairwise alignments and b progressive alignment. The four-letter designations are v-src, avian Rous sarcoma virus transforming factor; v-yes, avian Y73 sarcoma virus transforming factor; v-abl, Abelson murine leukemia virus transforming factor; v-fes, feline sarcoma virus transforming factor; v-fps, avian Fujinami virus transforming factor; v-raf, murine retroviral transforming factor; v-mos, mouse sarcoma virus transforming factor; cdc28, yeast cell division control factor; cadk, bovine cyclic AMP-dependent kinase. The same four-letter designations are used in Fig. 7 and Table 3.

```
                                                    * *  * *
v-src                        GLAK  DAW EIPRESLRLEAKLGQGCFGEVWMG    TWND T TRVAIKTLKPG TMSP EAFLQEA
v-yes                        GLAK  DAW EIPRESLRLEVKLGQGCFGEVWMG    TWNG T TKVAIKTLKLG TMMP EAFLQEA
v-abl                     TIYGVSPNYDKW EMERTDITMKHKLGGGQYGEVYEG    VWKKYS LTVAVKTLKED TMEV EEFLKEA
v-fes                      VLNRAVPKDKW VLNHEDLVLGEQIGRGNFGEVFSG    RLRADN TLVAVKSCRETLPPDIKAKFLQEA
v-fps                      VLTRAVLKDKW VLNHEDVLLGERIGRGNFGEVFSG    RLRADN TPVAVKSCRETLPPELKAKFLQEA
v-raf                        SSYY   W KMEASEVMLSTRIGSGSFGTVYKG    KWHGD  VAVKILKVVDPTPEQLQA FRNEV
v-mos                        GLPRRLAWFSIDWEQVCLMHRLGSGGFGSVYKA    TYHGVP VAIKQVNKCTEDLRASQRSFWAEL
cdc28                         MSGELANYKR       LEKVGEGTYGVVYKALDLRPGQGQRVVALKKIRLESEDEGVPSTAIREIS
cadk          LAKAKEDFLKKWENPAQNTAHLDQFER       IKTLGTGSFGRVMLVKHMETGNHYAMKILDKQKVVKLKQ  IEHTLNEKR

v-src         QV  MKKLRHEKLVQLYAV  VSEEPI YIVIEYMSKGSL    LDFLKGEM    GKYLRLPQLVDMAAQIASGMAYVERMNY
v-yes         QI  MKKLRHDKLVPLYAV  VSEEPI YIVTEFMTKGSL    LDFLKEGE    GKFLKLPQLVDMAAQIADGMAYIERMNY
v-abl         AV  MKEIKHPNLVQLLGVC TREPPF YIITEFMTYGNL    LDYLRECN    RQEVSAVVLLYMATQISSAMEYLEKKNF
v-fes         KI  LKQYSHPNIVRLIGVC TQKQPI YIVMELVQGGDF    LTFLRT E    GARLRMKTLLQMVGDAAAGMEYLESKCC
v-fps         RI  LKQCNHPNIVRLIGVC TQKQPI YIVMELVQGGDF    LSFLRS K    GPRLKMKKLIKMMENAAAGMEYLESKHC
v-raf         AV  LRKTRHVNILLFMG Y MTKDNL AIVTQWCEGSSL    YKHLHV Q    ETKFQMFQLIDIARQTAQGMDYLHAKNI
v-mos         NIAGLRHDNIVRVVAASTRTPEDSNSLGTIIMEFGGNVTLHQVIYDATRSPEPLSCRKQLSLGKCLKYSLDVVNGLLFLHSQSI
cdc28         LLKELKDDNIVRLYDI   VHSDAHKL YLVFEF LDLDLKRYM EGIPKDQPLG    ADIVK  KFMMQLCKGIAYCHSHRI
cadk          ILQAVNFPFLVKL EF   SFKDNSNL YMVMEY VP GGEMFS H LRRIGRFS    EPHAR  FYAAQIVLTFEYLHSLDL

                 *    *           ***                          ***            *    *
v-src         VHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQGAK FPIK WTAPEAA   LYGRFTIKSDVWSFGILLTELTTKGRV
v-yes         IHRDLRAANILVGDNLVCKIADFGLARLIEDNEYTARQGAK FPIK WTAPEAA   LYGRFTIKSDVWSFGILLTELVTKGRV
v-abl         IHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAK FPIK WTAPESL   AYNKFSIKSDVWAFGVLLWEIATYGMS
v-fes         IHRDLAARNCLVTEKNVLKISDFGMSREAADGIYAASGGLRQVPVK WTAPEAL   NYGRYSSESDVWSFGILLWETFSLGAS
v-fps         IHRDLAARNCLVTEKNTLKISDFGMSRQEEDGVYASTGGMKQIPVK WTAPEAL   NYGWYSSESDVWSFGILLWEAFSLGAV
v-raf         IHRDMKSNNIFLHEGLTVKIGDFGLATVKSRWSGSQQVEQPTGSVL WMAPEVIRMQDDNPFSFQSDVYSYGIVLYELMA GEL
v-mos         LHLDLKPANILISEQDVCKISDFGCSQKLQDLRGRQASPPHIGGTYTHQAPEILKGEIATP   KADIYSFGITLWQMTT REV
cdc28         LHRDLKPQNLLINKDGNLKLGDFGLARAF GVPLRAYTHEIVTLWY  RAPEVLLGGKQYS  TGVDTWSIGCIFAEMCN RKP
cadk          IYRDLKPENLLIDQQGYIQVTDFGFAKR   VKGRTWTL CGTPEY  LAPEIIL SKGYN  KAVDWWALGVLIYEMAAGYPP

v-src         PYPGMVN R  EVLDQVERG   YRM PCP PECPESLHDLMCQCWRKDPEERPTFKYL       QAQLLPACVLEVAE
v-yes         PYPGMVN R  EVLEQVERG   YRM PCP QGCPESLHELMKLCWKKDPDERPTFEYI       QSFLEDYFTAAEPSGY
v-abl         PYPGIDL S  QVYELLEKD   YRM ERP EGCPEKVYELMRACWQWNPSDRPSFAEI       HQAFETMFQESSIS
v-fes         PYPNLSN Q  QTREFVEKG   GRL PCP ELCPDAVFRLMEQCWAYEPGQRPSFSAI       YQELQSIRKRHR
v-fps         PYANLSN Q  QTREAIEQG   VRL EPP EQCPEDVYRLMQRCWEYDPHRRPSFGAV       HQDLIAIRKRHR
v-raf         PYAHINNRD  QIIFMVGRGYASPDL SRLYKNCPKAIKRLVADCVKKVKEERPLFPQILSSIELLQHSLPKINRSAPE
v-mos         PYS  GEPQ  YVQYAVVAYNLRPSLAGAVFTASLTG  KALQNIIQSCWEARGL   QRPSAELLHRDL KAFRGTLG
cdc28         IFSGDSEIDQIFKIFRVLGTPNEAIWPDIVYLPDFKPSFPQWRR KDLSQVVPSL   DPRGIDLLDKLLAYDPINRISA
cadk          FFADQP I QIYE  KIVSG       KVRFPSHFSSDLKDLLR  NLLQV  DL   TKRFGNLKDGVNDIKNHKWF
```

Fig. 7. Multiple alignment of seven tyrosine kinase-like oncogene sequences and those of yeast cdc28 and bovine heart cyclic AMP-dependent kinase as determined by progressive method. Asterisks denote locations where all nine residues are identical.

day structure without regard for how the structures came to be.

Global optimization methods are designed to accommodate a set of sequences in a multiple alignment that maximizes overall similarity. Three-dimensional extensions of the Needleman-Wunsch algorithm, for example, have been used to achieve such alignments (Jue et al. 1980; Murata et al. 1985), and Johnson and Doolittle (1986) have used the overlapping approach pioneered by Fitch (1966, 1970) to generate four-way and five-way alignments. Again, these alignments are made without regard to historical detail.

Historical alignments are based on the notion that divergent evolution is fundamentally binary in nature. Long ago Dayhoff et al. (1972), noting that matrix methods greatly foreshorten the more ancient branches in evolutionary trees, used a common-

Table 3. Percent identities calculated from binary (upper triangle) and progressive (lower triangle) alignment methods

	Tyrosine kinase-like sequences								
	v-src	v-yes	v-abl	v-fes	v-fps	v-raf	v-mos	cdc28	cadk
v-src		84	47	43	41	32	30	25	26
v-yes	84		49	43	41	35	30	25	26
v-abl	47	49		41	41	28	24	25	24
v-fes	41	43	41		79	30	26	25	26
v-fps	40	41	41	79		30	29	27	27
v-raf	30	31	26	29	30		27	23	23
v-mos	23	24	20	22	24	26		25	19
cdc28	18	19	16	16	17	21	25		26
cadk	19	18	17	15	16	18	20	26	

Table 4. Numbers of pairwise alignments required to construct a phylogenetic tree by a progressive method[a]

Number of sequences	Initial pairwise alignments	Additional iterative alignments	Total
3	3	2	5
4	6	4	10
5	10	6	16
6	15	8	23
7	21	10	31
8	28	12	40
9	36	14	50
10	45	16	61
11	55	18	73

[a] Values are minimal numbers for simple trees; compound trees need an additional alignment for each subcluster. Also, occasional negative segments in some trees will necessitate additional alignments

ancestor approach to alignment and tree building that was historical in principle. The character-based approach that they used was much clumsier than matrix methods, however, and eventually was abandoned. Subsequently, Holmquist (1979, p 939) drew attention to the fact that parsimony methods err significantly, "the magnitude of the error increasing with the distance of the nodal sequence from the present," and, more recently, Penny and Hendy (1986) have expounded on the theme that the minimal tree cannot be the historical tree.

It is obvious that methods based on mere global optimization will consistently underestimate evolutionary distances among the least related members of the set, striving as they do to achieve maximum alignment scores. The need is to throttle the tendency for optimization while preserving the notion of similar residues replacing one another. The progressive alignment procedure presented here appears to achieve that end. In its favor, the trees generated from these alignments appear to be in accord with biological expectations.

Superoxide Dismutase Relationships

The copper–zinc superoxide dismutase sequences have been the subject of much debate since the possibility was raised that the sequence found in the prokaryote *Photobacterium leiognathi* might be the result of a horizontal gene transfer from its ponyfish host (Martin and Fridovich 1981). Although solid evidence to the contrary was provided by Steffens et al. (1983), the notion has refused to go away (Bannister and Parker 1985). Our thinking about this matter is wholly in accord with that recently expressed by Leunissen and De Jong (1986): to wit, there is no basis for supposing anything other than a conventional history of events. Indeed, either of the evolutionary trees in Fig. 2 ought to dispel thoughts of a horizontal gene transfer for this gene, the photobacter position being entirely consistent with what would be expected for a typical prokaryotic–eukaryotic divergence. On the other hand, the tree made from pairwise alignments (2a) does have an unreasonble arrangement for the fruitfly and yeast, whereas the progressive tree is quite in line with conventional phylogeny.

It should be pointed out in passing that an apparent speed-up in the rate of copper–zinc superoxide dismutase evolution has occurred among the vertebrates (Lee et al. 1985). Thus, the apparent differences between mammalian and *Drosophila* sequences are much greater than would be expected on the basis of a comparison of the *Drosophila* and yeast sequences. The fact that there appears to have been a relaxation of selection pressures on the vertebrate superoxide dismutase should not affect the branching order, of course.

Hemoglobins and Myoglobins

The progressive alignment scheme also yields reasonable results when applied to distantly related globin sequences. In contrast to phylogenies employing a maximum parsimony method (Goodman et al. 1974), the progressive method roots the lamprey

and hagfish globins to the same branch as other vertebrate hemoglobins. Interestingly, an early study employing the common ancestor approach (Dayhoff and Eck 1968) also had the lamprey in this position. With regard to the relationship of animal and plant globins, the depth of the differences warrants a good deal of caution. Nonetheless, the recently published bacterial globin sequence (Wakabayashi et al. 1986) resembles the plant globins more than it does the animal globins, and it is not impossible that an unusual genetic event involving plants and symbiotic bacteria has occurred. A larger study encompassing all the known invertebrate and plant globin sequences may reveal more about the evolutionary connections of these proteins.

Concluding Remark

It is not our intention to reopen past skirmishing about the relative merits of strict parsimony methods and alternative treatments (Fitch 1981; Holmquist and Jukes 1981). Nor is it our aim merely to add one more comment to the enormous literature on the construction of evolutionary trees with sequence data (Tateno et al. 1982; Hogeweg and Hesper 1984; Penny and Hendy 1986, to name but a few). Rather, we simply offer a heuristic procedure for a computer-determined multiple alignment of related amino acid sequences that can be effected rapidly by objective criteria. Evolutionary trees drawn directly from these alignments appear to be very much in accord with biological expectations.

Acknowledgments. We acknowledge many helpful discussions with Mark Johnson and Marcella McClure; we are especially grateful to Mark Johnson for writing the programs BORD and TREEplot. This work was supported by NIH Grant GM-34434 and a grant from the American Cancer Society.

Note Added in Proof. During the period since the acceptance of this article we have applied the procedure in numerous settings, and, in some cases, the final alignment was slightly imperfect. The situation was remedied, however, by aligning each new sequence, or set of sequences, with an average sequence of all the sequences already aligned. This was accomplished by simply looking up the matrix value for every pair of residues at each position and averaging them. We are grateful to Steve Hanks for bringing the problem to our attention and to Mark Johnson for helping with the solution.

References

Bajaj M, Blundell T (1984) Evolution and the tertiary structure of proteins. Ann Rev Biophys Bioeng 13:453–492

Bannister JV, Parker MW (1985) The presence of a copper/zinc superoxide dismutase in the bacterium *Photobacterium leiognathi*: a likely case of gene transfer from eukaryotes to prokaryotes. Proc Natl Acad Sci USA 82:149–152

Cannon RE, White JA, Scandalios JG (1987) Cloning of cDNA for maize superoxide dismutase 2 (SOD2). Proc Natl Acad Sci USA 84:179–183

Dayhoff MO, Eck RV (1968) Atlas of protein sequence and structure 1967–1968. National Biomedical Research Foundation, Silver Spring MD, p 19

Dayhoff MO, Park CM, McLaughlin PJ (1972) Building a phylogenetic tree: cytochrome c. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington DC, pp 7–16

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model for evolutionary change. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington DC, pp 345–358

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Feng DF, Johnson MS, Doolittle RF (1985) Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol 21:112–125

Fitch WM (1966) An improved method of testing for evolutionary homology. J Mol Biol 16:9–16

Fitch WM (1970) Further improvements in the method of testing for evolutionary homology among proteins. J Mol Biol 49:1–14

Fitch WM (1977) On the problem of discovering the most parsimonious tree. Am Nat 111:223–257

Fitch WM (1981) The old REH theory remains unsatisfactory and the new REH theory is problematical—a reply to Holmquist and Jukes. J Mol Evol 18:60–67

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 15:279–284

Fredman ML (1984) Computing evolutionary similarity measures with length independent gap penalties. Bull Math Biol 46:553–566

Goodman M, Moore GW, Barnabas J, Matsuda G (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. J Mol Evol 3:1–48

Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20:175–186

Holmquist R (1979) The method of parsimony: an experimental test and theoretical analysis of the adequacy of molecular restoration studies. J Mol Biol 135:939–958

Holmquist R, Jukes T (1981) The current status of REH theory. Reply to an essay by Fitch. J Mol Evol 18:47–59

Hunt LT, Hurst-Calderone S, Dayhoff MO (1978) Globins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington DC, pp 229–249

Jabusch JR, Farb DL, Kerschensteiner DA, Deutsch HF (1980) Some sulfhydryl properties and primary structure of human superoxide dismutase. Biochemistry 19:2310–2316

Johansen JT, Overballe-Petersen C, Martin B, Hasemann B, Svendsen I (1979) The complete amino acid sequence of copper–zinc superoxide dismutase from *Saccharomyces cerevisiae*. Carlsberg Res Commun 44:201–217

Johnson MS, Doolittle RF (1986) A method for the simultaneous alignment of three or more amino acid sequences. J Mol Evol 23:267–278

Jue RA, Woodbury NW, Doolittle RF (1980) Sequence homologies among *E. coli* ribosomal proteins: evidence for evolutionarily related groupings and internal duplications. J Mol Evol 15:129–148

Kernighan BW, Ritchie DM (1978) The C programming language. Prentice-Hall, Englewood Cliffs NJ

Klotz LC, Blanken RL (1981) A practical method for calculating evolutionary trees from sequence data. J Theor Biol 91:261–272

Lee YM, Friedman DJ, Ayala FJ (1985) Superoxide dismutase: an evolutionary puzzle. Proc Natl Acad Sci USA 82:824–828

Leunissen JAM, De Jong WW (1986) Copper/zinc superoxide dismutase: how likely is gene transfer from ponyfish to *Photobacterium leiognathi*? J Mol Evol 23:250–258

Martin JP, Fridovich I (1981) Evidence for a natural gene transfer from the ponyfish to its bioluminescent bacterial symbiont *Photobacter leiognathi*. J Biol Chem 256:6080–6089

Moore GM, Goodman M, Barnabas J (1973) An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. J Theor Biol 38:423–457

Murata M, Richardson JS, Sussman JL (1985) Simultaneous comparison of three protein sequences. Proc Natl Acad Sci USA 82:3073–3077

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

Penny D, Hendy M (1986) Estimating the reliability of evolutionary trees. Mol Biol Evol 3:403–417

Rocha HA, Bannister WH, Bannister JV (1984) The amino-acid sequence of copper/zinc superoxide dismutase from swordfish liver. Eur J Biochem 145:477–484

Sankoff D, Cedergren RJ, McKay WM (1982) A strategy for sequence phylogeny research. Nucleic Acids Res 10:421–431

Sellers PH (1974) Evolutionary distances. SIAM J Appl Math 26:787–793

Steffens GJ, Bannister JV, Bannister WH, Flohe L, Gunzler WA, Kim S-MA, Otting F (1983) The primary structure of Cu–Zn superoxide dismutase from *Photobacterium leiognathi*: evidence for a separate evolution of Cu–Zn superoxide dismutase in bacteria. Hoppe-Seyler's Z Physiol Chem 364:675–690

Steinman HM, Naik VR, Abernathy JL, Hill RL (1974) Bovine erythrocyte superoxide dismutase J Biol Chem 249:7326–7338

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol 18:387–404

Wakabayashi S, Matsubara H, Webster DA (1986) Primary sequence of a dimeric bacterial hemoglobin from *Vitreoscilla*. Nature 322:481–483

Zelenik M, Rudloff V, Braunitzer G (1979) Die Aminosauresequenz des monmeren Hamoglobins von Lampetra fluviatilis. Hoppe-Seyler's Z Physiol Chem 360:1879–1894

Received December 11, 1986/Revised January 21, 1987