J.Mol.Evol.5, 1-24 (1975)

Phylogenies from Amino Acid Sequences Aligned with Gaps: The Problem of Gap Weighting

WALTER M. FITCH and KERRY T. YASUNOBU

Department of Physiological Chemistry,
University of Wisconsin Medical School, Madison

Department of Biochemistry-Biophysics,
University of Hawaii, Manoa Campus, Honolulu

Received May 3, 1974

Summary. The common but generally overlooked problem of how best to construct phylogenies from orthologous amino acid sequences, when their alignment requires the placement therein of gaps denoting insertions/ deletions in the evolutionary history of their genes since their common ancestor, has been studied. Three diverse methods were examined: 1. each missing residue in a gap is weighted as equivalent to the average number of minimum nucleotide replacements in known conjugate amino acid pairs of those same two sequences, which weight necessarily differs for each pair of sequences; 2. each missing residue in a gap is weighted as equivalent to a fixed number of nucleotide replacements; and 3. each gap, regardless of length, is weighted as equivalent to a fixed number of nucleotide replacements. For the flavodoxins, each method yielded a different best tree and suggests that the choice of method may be crucial. For the plant ferredoxins, all methods give results inconsistent with botanical classification and suggests the sequences may not all be orthologous. For the bacterial ferredoxins, the method was less germane than the actual weight used, five different best trees being obtained depending upon the weight. The best tree for all ferredoxins (prokaryotic plus eukaryotic) combined proved to be greatly dependent upon the gap locations with several reasonable alignments yielding different best trees. They also suggest that functional equivalence may well prove to be a poor guide to which residues have a common ancestral codon. The rubredoxin sequences show that a partial internal gene duplication occurred in the *Pseudomonas* line, probably very soon after its divergence from the other genera. Together, the results clearly indicate that the phylogenetic answer one gets may greatly depend upon how one treats the gaps but they fail to indicate what treatment may be best.

This results partly from the fact that the phylogenies of the taxa represented are not known with sufficient confidence to be sure when the procedures are performing best.

Key words. Phylogeny - Insertion-Deletion - Protein Sequences - Ferredoxins - Rubredoxins - Flavodoxins - Taxonomy.

The best way of determining the most appropriate phylogeny (dendrogram, bifurcating tree) that accounts for a set of observations in the face of missing information for several taxa has concerned numerical taxonomists for some time (see Sneath & Sokal, 1973). The amino acid sequences of orthologous[1] proteins have not generally been considered in this context. Nevertheless, the same problems frequently arise, even if they are generally ignored. There are, however, two special problems that are peculiar to amino acid sequence data. The first is that one can not always be sure what information is missing; the second is that the information is generally missing as a result of various biological processes that are sufficiently well known to permit one to make quantitative statements about the missing information.

The first peculiarity arises anywhere that two sequences differ and the more they differ, the more difficult it is to determine whether local differences arise from nucleotide replacements, deletions, insertions, recombinations, or frameshifts. If the sequences are of unequal length, deletions or insertions are necessarily involved, but their precise locations may not be clearly determinable. The problem is greater the more distantly related two sequences are and will be illustrated here by several of the iron-sulfur proteins, some of which bridge the prokaryote-eukaryote kingdoms.

The second peculiarity arises in several ways: 1. some residues of the sequence may not have been determined; 2. some may be known imprecisely such as ASX, GLX, SAC (Serine, Alanine or Cysteine); 3. some may not exist relative to another sequence because of insertions or deletions in one or both lineages since their most recent common ancestor; or 4. the sequence may be in error. These problems also will

[1] Homologous genes are of two varieties (Fitch & Margoliash, 1970). Those that diverge following the speciation of the taxa from which they derive and whose ancestral history has a one-to-one correspondence to the history of those taxa from which they derive are called orthologous (ortho = exact). Those that diverge following a gene duplication and continue to diverge in parallel even in the same line of descent are called paralogous. Alpha and beta hemoglobin are paralogous.

be illustrated and reasonable approaches to their solution suggested. More importantly, it will be clearly shown that the phylogenetic answers show a great sensitivity to both the data and the methods of treatment.

METHODS

Sequence Alignment. The procedure used for aligning two sequences was that of Fitch which first requires a general examination to demonstrate homology and the approximate location of any necessary gaps (Fitch, 1966a, 1970a) and then places those gaps so as to minimize the number of nucleotide replacements required to convert a gene coding for one sequence into one for the other in those positions where there are no gaps in either sequence (Fitch, 1969). The two step procedure is designed to prevent the excessive but improbable numbers of gaps that would be introduced if the only criterion were the maximization of amino acid identities. An alternative procedure acceptable for this task is that of Needleman & Wunsch (1970).

The procedures just described apply only to two sequences or two sets of sequences possessing the property that, within one set, the sequences already have an accepted alignment and thus have an identical length. There is no procedure that aligns three different amino acid sequences simultaneously. Fortunately, it is almost always clear in such cases that a specific pair should first be aligned to each other and that pair then aligned to the third.

The advantage of comparing two sets of sequences (e.g. several bacterial ferredoxins versus several plant ferredoxins) over two individual sequences, one from each set, is that more information goes into making the decision.

The disadvantage is the complexity of the statistics and the process involved and the additional computational time required. There is an alternative approach. Since the comparison of two sets of sequences has implicit in it the phylogenetic assumption that the taxa represented within either set are more closely related to each other than any pair taken from each set, we may as well make the assumption explicit and take advantage of it. This is accomplished by constructing (or accepting) a phylogeny for the taxa within a set and using the procedure of Fitch (1971) to construct the ancestral sequence (the procedure of Moore et al., 1973, can also be used). This has the advantage that only two sequences need to be compared but, more importantly, they are sequences much closer in time to their common ancestor and therefore more likely to permit the correct location of gaps. The ancestral sequences are in terms of nucleotides rather than amino acids but that is no hindrance since the basic principles of comparison remain the same. In fact it can be an advantage. For example, there are codons for serine and leucine that require only a single nucleotide change, thus making them potential candidates for occupying conjugate[2] positions, i.e. being homologues. If, however, within one set, serine has been replaced by glycine and, in the other set, leucine has been replaced by methionine, the most likely codons for serine and leucine are AGY and YUG respectively (Y = C or U) and they therefore differ in three rather than in only one nucleotide, thus making them less likely candidates to occupy conjugate positions. A more crude procedure, but frequently quite valuable nonetheless, is to assume that the amino acid most common among the descendents was the amino acid of the ancestor.

Distance Measurements. Most taxonomic procedures build their dendogram or tree from a matrix showing the extent of dissimilarity (or similarity) between all pairs of taxa. In our case, the data are amino acid sequences and the dissimilarity arises from mutational events, generally nucleotide replacements, and their number can be thought of as the genetic distance between two sequences. Our problem is how to modify the estimated genetic distance in order to admit genetic processes other than nucleotide replacements.

Let i_k and j_k be two amino acids. We define $MV(i_k, j_k)$ as their value, i.e. the minimum number of nucleotides that must be changed to convert the coding from one amino acid to the other. Examples, using the IUB single letter code, are $MV(M,M) = 0$; $MV(M,L) = 1$; $MV(M,P) = 2$; $MV(M,H) = 3$.

Assume we have two homologous amino acid sequences, so aligned (with gaps, if necessary) that the m^th amino acids of the two sequences are conjugate (i.e., they are presumed to share a common ancestor). We define d_k as the distance (minimum nucleotide differences) over pairs of known amino acids and, by inference, therefore excluding pairs where one or both residues are either unknown (X) or a gap (*), e.g.

[2] Two sequences possessing a common ancestor are said to be homologous and two amino acids whose codons have a common ancestor might also be called homologous. Problems associated with a clear definition of varying types of genetic recombination (see Fitch, 1973) require an alternative word. Such amino acids have been called conjugate (after the linguistic usage that describes the common origin of words such as dale and dell) and that term will be used here.

$MV(A,X) = 0$ and $MV(A,*) = 0$. Then

$$d_k = \sum_{m=1}^{s} MV(i_m, j_m)$$

where i and j are the two sequences and s is their length.

We define r as the replacement rate per codon and k as the number of codon positions for which the amino acids are known in both sequences. Hence $k \leq s$ and $r = d_k/k$.

We define a *common* gap between two sequences as one in which an uninterrupted series of gap residues begin and end at the same residue positions in both sequences. Assuming that the most reasonable explanation of a common gap is an insertion or deletion *prior to* the common ancestor that diverged to give rise to the two sequences being examined, then that insertion or deletion should not count as an insertion or deletion in determining the distance (or change) since their common ancestor. Instead we treat it as so many residues that, had they been present, would have diverged to approximately the same extent as the remaining amino acids, i.e. these residue positions are treated as if they were X's that would have diverged at the rate r.

We further define n as the number of positions where one or both residues of the pair is a gap not in common. We also define u as the number of positions where one or both residues of the pair exists but is unknown (X), but neither one is a gap, or where both are members of a *common* gap. Thus, $s = k + n + u$.

We define d_a as the minimum distance adjusted for unknown amino acids and assume that if we knew the amino acids present in these positions, they would have the same replacement rate as in those positions where we know both amino acids. Thus we get a base distance, for all positions except those with non-common gaps, of $d_a = d_k + ru$.

We can make three general assumptions regarding the mutation value of gaps. The first is that the mutation value or effective replacement rate should be equivalent to that in the known positions[3]. This is probably too low a value

[3] The implicit assumption for such a computation is that the deleted or inserted positions be representative of all positions with respect to their rate of nucleotide substitution. An alternative treatment would consider the sequence A to be divided into two regions, the deleted and the undeleted regions. For all sequences still retaining the deleted region of A, determine the rate of change in the deleted region relative to that in the undeleted region and assume this relative rate would have applied to the deleted portion of sequence A had it not been deleted.

since insertions and deletions are quite rare compared to nucleotide replacements and should therefore be more heavily weighted. Nevertheless, letting d_u be the distance adjusted as if all gap residues were unknown amino acids, $d_u=d_a+rn=rs$.

This is the equation used by Fitch & Margoliash (1967). It, most clearly, had been used by Matsubara et al. (1968) for their ferredoxin study, recognizing of course that d_u/s = MBDC in their terminology.

An alternative assumption is that each residue in a non-common gap should be treated as equivalent to a mutation value of m_n nucleotide replacements. One possibility is that m_n be set equal to three since three nucleotides have been deleted, but any value is in principle acceptable. We define d_n as the minimum distance adjusted for the number of residues in non-common gaps. Then, $d_n = d_a + m_n n$.

Still a third alternative assumption stresses the fact that the insertion or deletion was the genetic event and that we should evaluate it on the basis of the number of gaps rather than their total length. In such a case we may treat each non-common gap, regardless of length, as equivalent to a mutation value of m_g nucleotide replacements. If we let g equal the number of gaps and d_g equal the minimum distance adjusted for the number of non-common gaps, then $d_g = d_a + m_g g$.

If there are no X's, $d_a = d_k$. If there are no gaps, $d_u = d_n = d_g = d_a$. Note that two non-common gaps must be counted as two events even if one spans the other.

The d_n computations could easily lead to some absurdities when two sequences of vastly different length are compared. For example, if the 55 amino acids of a bacterial ferredoxin were aligned with the first 55 of a plant ferredoxin, there would remain a sequence of 42 unmatched plant ferredoxin amino acids which would require a correspondingly long terminal gap for the bacterial sequence. Any correction based upon the number of residues in gaps becomes dominated by the huge extra piece on the end. Since terminal deletions are more common than internal deletions, the implied weighting of terminal deletions by gap length is particularly inappropriate. Therefore a rule was adopted as follows. Only one residue of a sequence (or of a set of sequences of common length in a given direction) shall be considered beyond that of the last considered residue of the sequence(s) extending next most in that direction. In such cases, at least one sequence should have but a single * at the appropriate end. This implies that no sequence begins or ends in such a fashion as to imply that two or more sequences all begin or end at a particular residue when only part of them do.

Constructing Phylogenies. The procedure of Fitch & Margoliash (1967) was used for this purpose except for one modification; namely a weighting of distances according to the principle that each side of a bifurcation should receive equal weight. Thus, if each side gives rise to the same number of taxa, the distances to these various taxa are equally weighted but if one side gives rise to a single taxon and the other side gives rise to two taxa, distances to the two taxa are each weighted one half of that to the taxon on the other side. This is appropriate because we are unwilling to assume that both sides evolved at the same rate which is the only condition under which the distance from an outside taxon to the lone taxon is no more representative of that distance than all the other distances to the taxa on the other of the bifurcation. We are saying that multiple taxa on one side are expected to increase the accuracy of the estimated distance on that side by reducing the variance of the estimate, but, in the absence of the uniform rate assumption, they contribute nothing to the estimate of the distance on the other side of the bifurcation, and each side has an otherwise equal claim to be an estimate of the distance to the outside taxa. Indeed they must be so separated if the rates in different lines of descent are to have the maximum opportunity to differ.

RESULTS AND DISCUSSION

Table 1 shows the pairwise distances between the four known flavodoxins, whose alignment is shown in Fig.1. Below the

Table 1. Pairwise distances for four flavodoxins

		1	2	3	4
1	*Cl. MP*	-	113.1	121.8	158.0
2	*Cl. pasteurianum*	108.3	-	119.6	155.5
3	*Pe. elsdenii*	108.9	107.1	-	167.0
4	*D. vulgaris*	151.1	151.6	162.2	-

The values below the principle diagonal are the original distances adjusted for all unknown amino acids (d_u; see text). Those above the diagonal are the original distances adjusted for the number of non-common gaps (d_g). The mutation value for a gap, m_g, was set at 3. The calculations assumed the alignment shown in Fig.1. The phylogenies most appropriate to these two data sets are shown in Fig.2. The references for the sequences in the order given above are: 1, Tanaka et al. (1974a); 2, Tanaka et al. (1971b); Fox et al. (1972); 3, Tanaka et al. (1971b, 1974a); 4, Dubourdieu et al. (1973). The genus abbreviations are: *Cl., Clostridium; Pe., Peptostreptococcus; and D., Desulfovibrio.*

```
                       10                  20                  30                  40
m.    M * K * * I V Y W S G T G N T E K M A E L I A K G I I E S G K D V N T I N V S D
p.    M * K V N I I Y W S G T G N T E A M A K L I A E G A Q E K G A Q V K L L N V S D
e.    M * * V E I V Y W S G T G N T E A M A N E I E A A V K A A G A D V E S V R F E D
v.    M P K A L I V Y G S T T G N T E Y T A E T I A R E L A B A G Y E V D S R O A A S

                       50                  60                  70                  80
m.    V N I D E L L N E * D I L I L G C S A M G D E V L E E S E F E P F I E E I S T K
p.    A K E D D V K E A * D V V A F G S P S M G S E V S E * * M Z Z P F L D V V S S I
e.    T N V D D V A S K * D V I L L G C P A M G S E E L E D S V V E P F F T D L A P K
v.    V E A G G L F E G F D L V L L G C S T W G D D S I Q * * L Z B B F I P L F D S L

                       90                 100                 110                 120
m.    * * * * I S G K K V A L F G S Y G W G D G K W M R D F E E R M N G Y G C V V V E
p.    * * * * V T G K K E G A F X X X X X X X X X X X X X X X X X X X X X X X X X X X
e.    * * * * L K G K K V G L F G S Y G W G S G E W M D A W K Q R T E D T G A T V I G
v.    Z Z T G A Z G R K V A C F G C * * * G B S S Y E Y F C G A V D A I E E K L K N L

                      130                 140
m.    T P L I V Q N E P D E A E Q D C I E F G K K I A N I *
p.    X X X X X X X X X X X X X X X X X N I G R E L V * * *
e.    T A * I V N E M P D N A P E * C K E L G E A A A K A *
v.    G A Z I V Z B G L R I D G D P R A A R B B I V G W A H
```

Fig.1. Alignment of flavodoxins. The amino acid sequences are given in the IUB single letter code plus X for an undetermined amino acid and an * for gaps inserted to optimize the correspondence between the sequences. The sequences derive from the following taxa: *m., Clostridium MP; p. Clostridium pasteurianum; e., Peptostreptococcus elsdenii; and v., Desulfovibrio vulgaris*. The number of residues between the two partial sequences of *pasteurianum* ferredoxin is not known and hence could conceivably contain positions that should have been denoted by an *

principle diagonal are the d_u distances in which all residue pairs involving either an unknown residue or a gap are given a mutation value equal to the average number of nucleotide differences per codon (r) for which both amino acids are known in the same two sequences. Above the diagonal are d_g distances in which every gap, regardless of length, is assigned a mutation value per gap, m_g, of 3 instead of each individual residue being assigned the r value. The result of this change is that the best tree for the d_g data is different from that for the d_u data. These two trees are shown in Fig.2, where it can be seen that it requires the gap adjustment to cluster the two *Clostridia* together. This tree arises for all m_g>1.3. One also alters the result if the

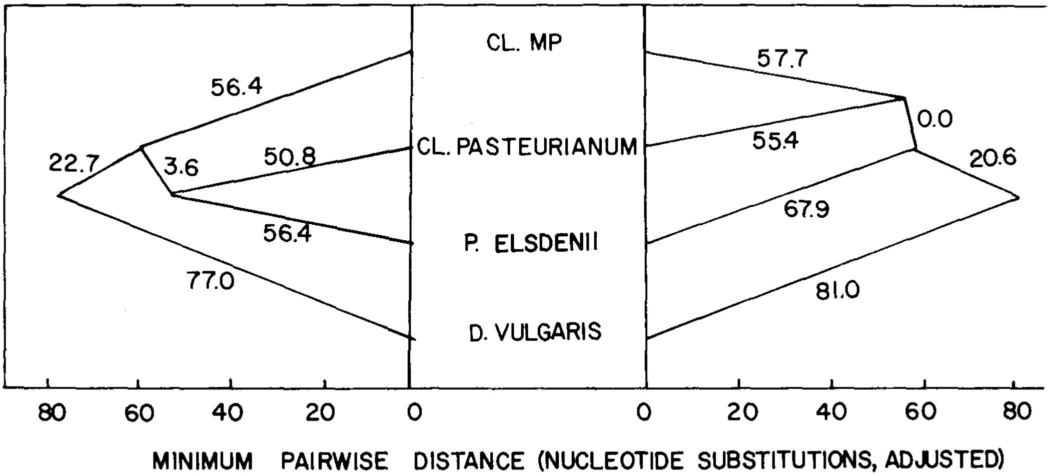

Fig.2. Flavodoxin phylogenies. The left-hand tree is based on the d_u distances in Table 1, the right-hand tree on the d_g distances

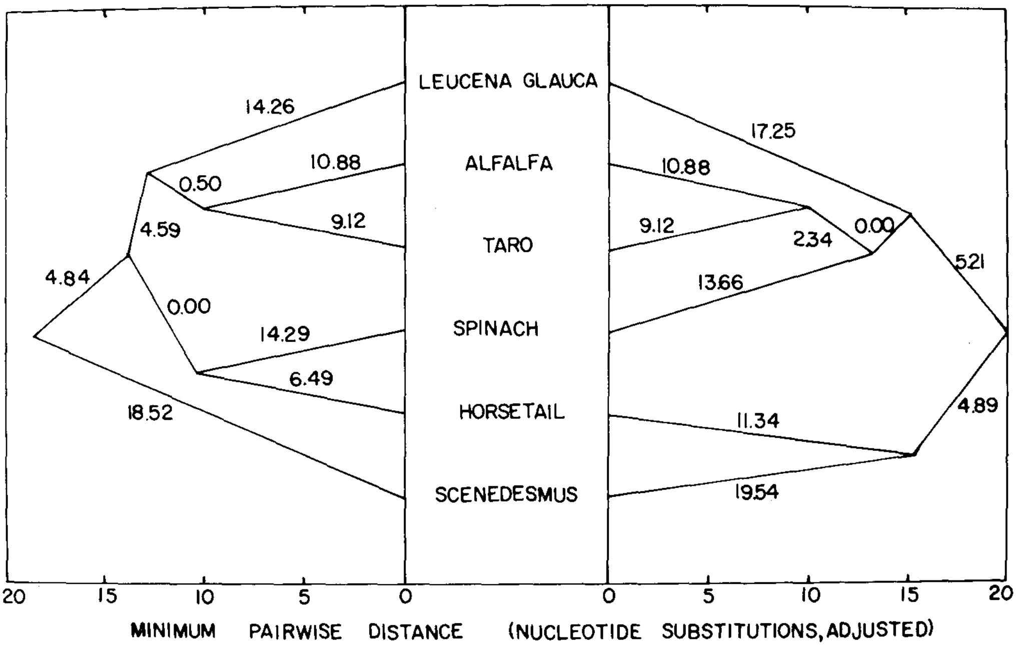

Fig.3. Plant ferredoxin phylogenies. The left-hand tree is based on the d_u distances in Table 2, the right-hand tree on the d_n distances

distance is adjusted on the basis of the length of the gap, but this time a third tree is obtained which forms the first node by clustering *Cl. MP* with *P.elsdenii* for all m_n>1.1. This result is as bad as that for d_u distances if the *Clostridia* are of monophyletic origin and *Peptostreptococcus* belongs in·a separate genus.

Table 2. Pairwise distances for six plant ferredoxins

	1	2	3	4	5	6
1 Taro	-	20.00	27.00	25.00	35.86	39.00
2 Alfalfa	20.00	-	28.00	27.00	35.86	43.00
3 *L.glauca*	24.25	25.26	-	33.00	33.62	50.00
4 Spinach	25.00	27.00	30.31	-	29.14	42.00
5 Horsetail	27.71	27.71	25.40	20.78	-	30.88
6 *Scenedesmus*	36.38	40.42	44.93	39.41	25.40	-

The values below the principle diagonal are the original distances adjusted for all unknown amino acids (d_u; see text). Those above the diagonal are the original distances adjusted for the number of residues in non-common gaps (d_n). The mutation value for a gap residue, m_n, was set at 3. The phylogenies most appropriate to these two data sets are shown in Fig.3. The references for the sequences in the order given are: 1, *Colocasia esculenta*, Rao & Matsubara (1970); 2, *Medicago sativa*, Keresztes-Nagy et al. (1969); 3, *Leucaena glauca*, Benson & Yasunobu (1969); 4, *Spinacea oleracea,* Matsubara et al. (1967, 1968); 5, Aggarwal et al. (1971); and 6, Sugeno & Matsubara (1969).

The alignment used is identical to that of the Atlas (Dayhoff, 1972). The spinach sequence is the one containing glutamate and isoleucine in positions 31 and 33, respectively.

This is an excellent illustrative example in that it shows three different best trees depending upon whether one corrects for the number of gaps not in common, or for the length of gaps not in common, or whether one simply treats all unknown residue pairs as equivalent to the average of the known pairs. Unfortunately, one cannot say anything about the relative merits of the three approaches because of the extensive unsequenced region of the *Cl.pasteurianum* flavodoxin. Even the length of that region is uncertain and it might be noted that the C-terminal heptapeptide would require only four nucleotide changes for its codons to be able, potentially, to encode the heptapeptide beginning 26 residues back at residue 116 of the *D.vulgaris* flavodoxin. Moreover, 13 of the *D.gigas* residues are B or Z, i.e. their amidation state is unknown. The phylogenetic results are very nearly identical if the heptapeptide is paired instead with the last seven of *D.vulgaris* instead of back two residues. This would however increase all pairwise distances to *D.gigas*. These results strongly suggest that sequences with considerable numbers of uncertain residues ought not to be included if phylogenetic relationships are sought.

Six plant ferredoxins, including that from horsetail (*Equisetum*) have been similarly examined. There are only three gaps, an N-terminal gap of one in *L.glauca* ferredoxin, an N-terminal gap of two in horsetail and a common C-terminal gap of one in horsetail and *Scenedesmus*. Nevertheless, the mutation value placed on the insertion/deletion process affects the topology of the best fitting tree. Table 2 shows the pairwise distances d_u and $d_n(m_n = 3)$. Fig.3 shows the two best trees for these two sets of distances. Compared to botanical expectations, neither tree is very satisfactory. In one tree, spinach is more closely related to horsetail than to the other dicots and, in the other tree, horsetail is more closely related to the algal *Scenedesmus* than to other vascular plants. Moreover, in both trees, alfalfa is more closely related to the monocot, taro, than to the other legume, *L.glauca*. The horsetail amonalies might be attributable to the fact that it is only half sequenced and the variability in other sequences appears to be somewhat greater in the regions for which horsetail ferredoxin is not sequenced compared to those regions where it is. This suggests that partial sequences should not lightly be included among complete sequences.

The difficulty with the taro relationship may simply reflect widely different rates of evolution in various lines of descent, but there is another possibility, already observed once before, that should receive serious consideration, namely, paralogous genes[4]. The lysozyme sequences of chicken and duck were vastly more similar than either was to the goose lysozyme sequence and led to unrealistic conclusions (duck more closely related to the chicken than to the goose or an enormous evolutionary rate for goose lysozyme since the common ancestor of duck and goose). This problem was resolved when Arnheim & Steller (1970) discovered that the black swan has two lysozymes, one similar to the goose, the other to the duck-chicken lysozyme and that birds seem generally to possess both forms but express them at different stages of their development. One might therefore be legitimately concerned that some such event is occurring among the ferredoxins. For example, the spinach and *L.glauca* ferredoxin sequences might be from a second more rapidly evolving locus than the locus producing the alfalfa and taro sequences.

Eight anaerobic bacterial ferredoxins have been examined. Their pairwise d_u are shown in Table 3 and the best tree for these distances is shown in Fig. 4. The pairwise distances reconstructed from Fig.4 are also shown in Table 3. *Peptococcus*

[4] see footnote 1.

Table 3. Pairwise distances for eight bacterial ferredoxins

	1	2	3	4	5	6	7	8
1 *Cl.butyricum*	-	13.24	20.34	21.66	27.75	30.32	32.36	41.49
2 *Cl.pasteurianum*	13.24	-	18.56	19.88	25.97	28.54	30.58	40.16
3 *Cl.acidi-urici*	21.38	18.33	-	19.02	25.11	27.68	29.72	39.30
4 *P.aerogenes*	20.08	20.08	19.02	-	26.43	29.00	31.04	40.62
5 *Cl.ME*	28.51	23.42	23.42	29.58	-	31.75	33.79	43.37
6 *Cl.thermosaccharolyticum*	33.60	32.58	29.53	28.53	30.54	-	2.04	25.94
7 *Cl.tartarivorum*	35.64	34.62	31.56	30.64	32.58	2.04	-	27.98
8 *Pe.elsdenii*	37.33	40.44	38.37	36.63	44.59	25.92	28.00	-

The values below the principle diagonal are the original distances adjusted for all unknown amino acids (d_u; see text). Those above the diagonal are the reconstructed distances obtained by summing the values on the legs in Fig.4. The references for the sequences, in the order given above, are: 1, Benson et al. (1967); 2, Tanaka et al. (1966); 3, Rall et al. (1969); 4, Tsunoda et al. (1968); 5, Tanaka et al. (1974b); 6, Tanaka et al. (1973a); 7, Tanaka et al. (1971a); and 8, Azari et al. (1970) and Yasunobu & Tanaka (1973a). The alignment used is identical to that of the Atlas (Dayhoff, 1972). The genus abbreviations are: *Cl.,Clostridium; P.,Peptococcus; and Pe.,Peptostreptococcus.*

(previously called *Micrococcus*) has two gaps of one residue each plus one extra C-terminal residue. *Peptostreptococcus* has one gap of one residue. Since all gaps are of length one, it is immaterial whether one corrects for the number of gaps or the length of the gaps. The effect of increasing the value of *m* is startling however. Five different best trees are obtained for *m* values between zero and six and are shown in Fig.5. Moreover, if no correction for gaps is made, *Peptococcus* appears closely related to *Cl.acidi-urici* and *Peptostreptococcus* appears closely related to the heat stable *Cl. tartarivoran* and *Cl.thermosaccharolyticum*. However, as *m* is increased, the relatedness decreases and the *Clostridia* become more and more clustered until, for values of $m>5.62$, the *Clostridia* finally emerge as a monophyletic group. If nothing else, this dramatizes the importance of making a proper allowance for the effect of gaps on the distance measures.

In the preceding cases, the proper location of gaps was hardly in doubt. The same cannot be said if one attempts to align the bacterial and plant ferredoxins along with those

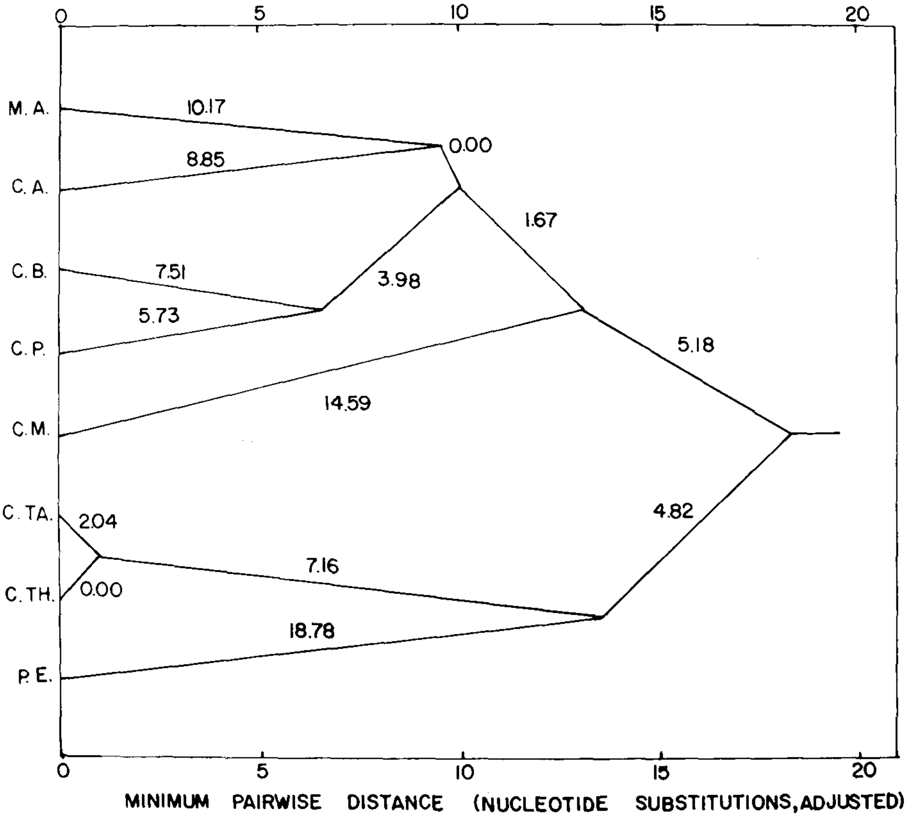

Fig.4. Bacterial ferredoxin phylogenies. Tree is based on the d_u distances in Table 3

of *Chlorobium, Chromatium* and *Desulfovibrio*. Several possibilities, all of which differ from that of Dayhoff (1972) and in particular with her alignment of the plant ferredoxins, are shown in Fig.6. Distances for each of the three forms, d_u, d_n and d_g, are given in Table 4 for values of m_n and m_g equal to three and five respectively. The best tree, as shown in Fig.8, is dependent upon both the alignment and the value of *m*. On what grounds then, might one prefer one alignment over another?

Measures that might be used to judge the relative merits of two alternative alignments for a set of sequences include: 1. the shortest overall length of sequences; 2. the lowest minimum replacements per codon, $\bar{r}$ (the average *r* over all pairs of sequences); 3. the lowest average distance $\bar{d}$ $(=\bar{r}s)$[5];

[5] It is obvious that *r* can be continually reduced by increasing the number of gaps but this is paid for by increasing the sequence length *s*. For real data, the introduction of a few gaps will frequently reduce $rs = d_u$ but d_u reaches a minimum following which it rises. It might

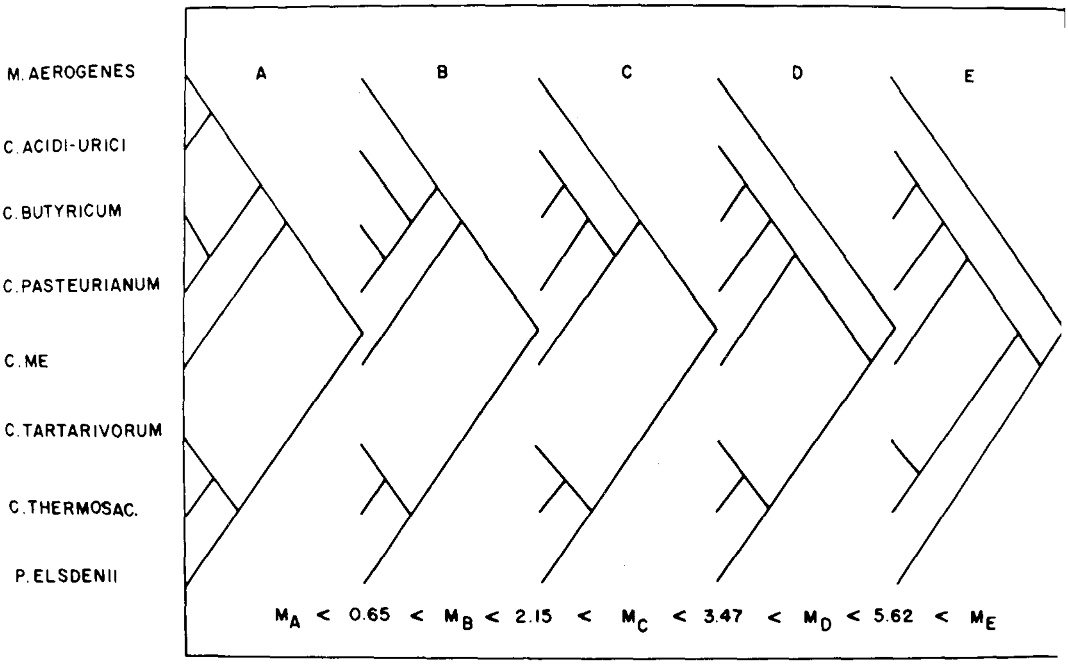

Fig.5. Bacterial ferredoxin phylogenies. Each tree is representative only of the branching order, branch lengths having no significance. Each topology is the best tree for a different set of pairwise distances that differ only as a function of the mutation values assigned per gap (m_g) and per gap residue (m_n). Since all gaps are of length one, m_g and m_n are indistinguishable. The range of m values that give rise to a particular topology is given below that tree. The complete range of all m is bounded by zero and infinity

4. the fewest gaps; 5. the greatest consistency among the tree distances, d_u, d_n and d_g; and 6. the least "information" required to interconvert the genes of the various proteins (Reichert et al., 1974). The last is another form of distance but has probabilistic roots in thermodynamics and, although it might ultimately be the best, our attention is here is restricted to the first five.

therefore be thought that a criterion for gap limitation is the point at which d_u is minimized. Unfortunately, it is easily demonstrated that a given set of potential gap placements for a given portion of a sequence may be acceptable or not depending solely on the total length of the sequence (holding r constant) or, holding s constant, depending solely upon the value of r. The larger the values of r or s, the more likely a given gapping proposal will be acceptable.

```
                                 1                 10                   20
Chlorobium limicola              A L Y I T E E C T Y C G A C E P E C P V T
Composite Anaerobic Bacterium    A H V I N D E C I S C G A C A X E C P V X
Chromatium                       A L M I T D Q C I N C N V C Q P E C P N G
Desulfovibrio gigas              * P I Q V D N C M A C Q A C I N E C P V D
Composite Plant                  A A Y K V K L V T P X G X Q E F E C P D D

22              30                40                   50                   60    6
A I S * * * A G D D I Y V I D * A N T C N E * * C A G L B Z Z A * * * * C V * * *
A I X * * * Z G D S K Y V I D * A D T C I D * * C G A * * * * * * * * * C A * * *
A I S * * * Q G D E I Y V I E * P S L C T E * * C V G H Y E T S Q C V D C V * * *
V F Q M D E Q G D * K * A V N I P N S N L D D Q C V E * * * * * * * * * A I * * *
V|Y I L D * Q A E E X * G I D L|P Y S C R A|G S C S S * * * * * * * * * C A G K V|

  A|* * I S A G D D I Y V I D|A                 E|C A G L B Z * * * * Z A C V|A
  A * * I X Z G D S K Y V I D A                 D * * * * * * * * * C G A C A A
  A * * I S Q G D E I Y V I E P                 E C V G H Y E T S Q C V D C V E
  V F Q M D E Q G D K A V N I P                 D * * * * * * * D Q C V E A I Q
  V|Y I L D Q A E E X G I D L|P                 A|G S C S S * * * * C A C K V|X

      Alternative A                                  Alternative B
                           64         70    74
                           V C P A E C I V Q G *
                           V C P V G A P X Q E *
                           V C P I K D P S H E E
                           S C P A A I R S * * *
                           X G X V D Q S D Z S F
```

Fig.6. Alignments of five diverse ferredoxins sequences. Representation is as in Fig.1. Alignment 1 is the one shown. Alignment 2 substitutes alternative A for the region between residues 22 and 38. Alignment 3 substitutes alternative B for the region between residues 43 and 63. Alignment 4 substitutes both alternatives A and B. The APS alignment is from the Atlas of Protein Sequences (Dayhoff, 1972) adjusted only as necessary to include *D.gigas* but not altering the relationships of the other sequences to each other. This amounted to a single gap of two residues opposite the Phe-Gln sequence in *D.gigas*. Also an extra column of * at the beginning and end of the prokaryotic sequences was required because only the plant sequences extended without interruption beyond those residues. The *Chromatium* sequence extends beyond that shown in alignment 1-4 but may be considered to continue its alignment to the plant without interruption. This is shown in Fig.7 (*Chromatium* 1). The composite bacterial and plant sequences were formed by inspection of the respective sequences (basically, majority rules with X's used where choice appeared uncertain)

Table 4. Pairwise distances for five diverse ferredoxins

	Minimum pairwise nucleotide differences adjusted for														
	All unknown residues (d_u)					Number of residues in gaps (d_n)					Number of gaps (d_g)				
Alignment	1	2	3	4	APS	1	2	3	4	APS	1	2	3	4	APS
Chlorobium- Bacteria	36.3	35.3	33.1	32.2	34.8	58.9	57.9	58.3	57.4	57.4	41.9	40.9	38.3	37.4	45.4
Chlorobium- *Chromatium*	51.8	50.4	47.2	45.8	48.5	63.3	61.9	58.7	57.4	62.4	58.3	56.9	53.7	52.4	59.4
Bacteria- *Chromatium*	46.4	45.2	43.3	42.0	44.5	70.2	68.9	67.0	65.8	70.6	50.2	48.9	47.0	45.8	57.6
Chlorobium- *Desulfovibrio*	78.4	78.9	74.6	74.9	80.6	119.2	111.3	113.0	105.0	112.3	106.2	95.3	98.0	87.0	96.3
Bacteria- *Desulfovibrio*	64.3	70.5	60.0	65.6	72.5	89.9	86.7	100.5	95.9	110.1	93.9	87.7	88.5	80.9	93.1
Chromatium- *Desulfovibrio*	75.5	78.9	70.3	73.2	77.6	117.1	111.3	104.0	98.0	106.2	94.1	85.3	88.0	79.0	91.2
Chlorobium- Plants	74.0	79.3	73.0	77.1	79.5	114.0	111.6	94.3	91.9	107.7	99.0	90.6	96.3	82.9	107.7
Bacteria- Plants	70.6	75.2	72.1	75.9	67.8	93.1	90.8	101.5	98.3	100.5	95.1	86.8	91.5	82.3	92.5
Chromatium- Plants	86.1	89.3	83.1	85.3	80.6	121.0	117.4	99.2	95.6	104.8	99.0	89.4	97.2	87.6	105.8
Desulfovibrio- Plants	84.6	85.0	73.4	74.0	92.5	101.3	97.7	102.4	98.6	133.2	99.3	91.7	87.4	79.6	111.2

Calculations were performed as given in the text. The values of m_n and m_g were 3.0 and 5.0 respectively. The alignments are given in Fig.6 except for APS which is from the Atlas of Protein Sequences (Dayhoff, 1972). References for the above sequences are: *Chlorobium limicola*, Tanaka et al. (1974c); Bacteria, see Table 3; *Chromatium*, Matsubara et al. (1970); *Desulfovibrio gigas*, Travis et al. (1971); Plants, see Table 2

```
                        75        80                  90                  100             108
                                    D
Composite Plant         L D D X Q I   E G W V L T C V A Y P X S D V T I E T H K E E E L T A ·
                                    E
            1           T E D E L R A K Y E R I T G E G ·
Chromatium  2                     T E D E L               R A K Y E R I       T G E G ·
            3                                 T E D E L R A K Y E R I T G E G ·
            4                                                     T E D E L R A →
```

Fig.7. Alignments of C-terminus of plant and chromatium ferredoxins. Shown are four alternatives. All 4 alignments require a terminal deletion. In addition number 1 differs from the plant sequence shown by a minimum of 17 nucleotides, number 2 differs by 13 nucleotides and 3 internal deletions, number 3 by 17 nucleotides and 1 internal deletion, and number 4 by 3 nucleotides, 17 unmatched *Chromatium* nucleotides and 1 internal deletion. Number 4 is that in the Atlas (Dayhoff, 1972). A · signifies the C-terminus, a → signifies that the sequences continues

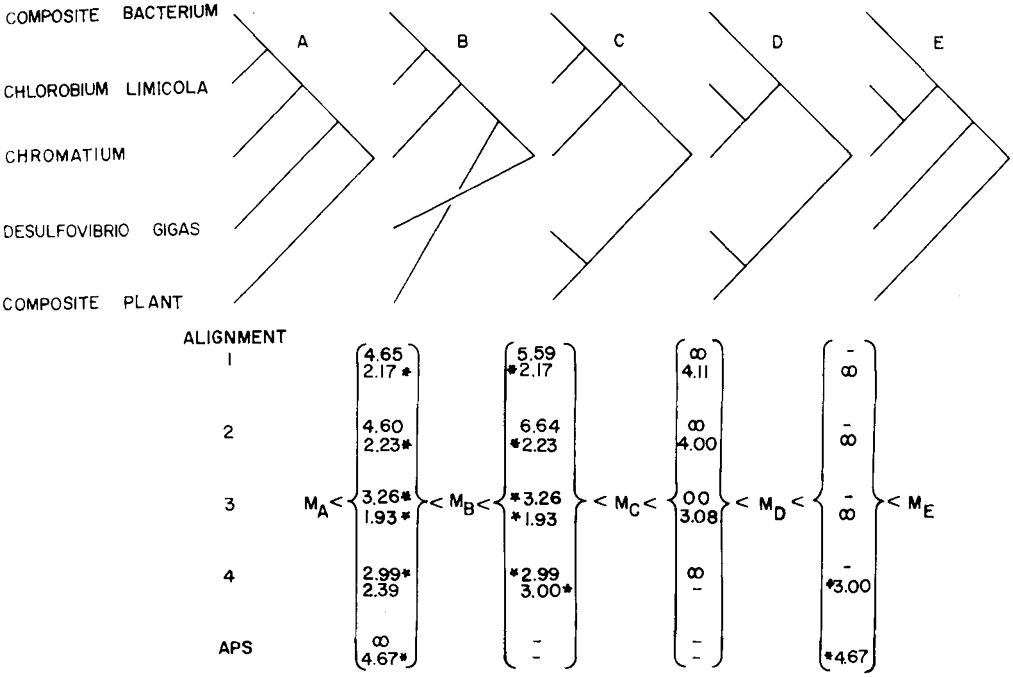

Fig.8. Phylogenies for five diverse ferredoxins. Each tree is representative only of the branching order, branch lengths having no significance. Each topology is the best tree for a different set of pairwise distances and sequence alignments. The range of *m* values that lead to each topology is given in pairs for each alignment. The upper member of each pair is a bound for m_g, the lower member a bound for m_n. The * symbols occur in pairs that separate identical bounds of *m* and indicate that any topology above that physical interval is not a best tree for any value of *m*. The - symbol indicates no bound is required. The alignments are defined in Fig.6

Table 5. Comparison of five diverse ferredoxin alignments

Alignment	s	$\bar{r}$	$\bar{d}_u$	gaps
1	74	.896	66.3	12
2	72	.937	67.5	8
3	69	.909	62.7	12
4	67	.961	64.4	8
APS	71	.946	67.2	15

The alignments are those given in Fig.6. The symbols are: s, the length of the five aligned sequences; $\bar{r}$, the average replacements per codon found by summing the total number of nucleotide differences for all ten pairwise sequences and dividing the total number of positions in those ten comparisons where neither member of a pair of amino acids was represented as being unknown; $\bar{d}_u = \bar{r}s$; gaps is the number of distinct gaps in the five sequences, distinct meaning that common gaps are counted only once.

The first four of these are shown in Table 5, where it can be seen that alignment 1 has the lowest r, alignment 2 (and 4) has the lowest number of gaps, alignment 3 has the lowest average distance, and alignment 4 has the shortest sequence. The APS (Atlas of Protein Sequence, Dayhoff, 1972) alignment seems poor by all four standards. However the APS alignment is clearly superior on the basis of the consistency of the three distance measures since the tree that one obtains for d_u is the same tree as that for d_g for all values of m_g and the same as for d_n for all values of $m_n<4.67$. Moreover three relatively closely related sequences (*Chlorobium, Chromatium,* and composite bacteria) all have very similar distances to *D.gigas* or to composite plants, especially when compared to the other four alignments. To the extent that r, g (number of gaps) and n (number of gapped residues) all increase monotonically as time since divergence increases, this consistency is what one should expect. However, if that relationship truly held, correcting for gaps would be unnecessary. In any event, we have five alignments, each one ostensibly best by a different one of five criteria.

The preceding criteria are internal. We can also compare the resulting trees to those expected on other grounds. We can see from Fig.8 that only trees A and E preserve the separateness of the eukaryotic plant ferredoxin from the other four prokaryotic sequences. Between the two trees A and E, the choice is only whether to link *Chlorobium* more closely to the composite anaerobic bacteria (A) or to *Chroma-*

```
Chlorobium       E C A G L B Z * Z A C V * * * A
Bacteria         D C G * * * * * * A C A * * * A
Chromatium       E C V G H Y E T S Q C V D C V E
Desulfovibrio    D D Q * * * * * * * C V E A I Q
Plants           A G S C S S * * * * C A G K V X

                        Alternative C
```

Fig.9. Alternative alignment for the region of five diverse ferredoxins. This alternative, C, is an alternative to B in Fig.6

tium (E). As can be seen, no value of m_g gives the E result and only values of m_n>3.08 would, in the most favorable case (alignment 3), give the E result.

Thus, tree A seems clearly to be the phylogeny of choice using these sequences. That unfortunately tells us little regarding the most appropriate alignment since all five alignments yield tree A when m = 0. Moreover, one expects that almost all alignments will give the same tree for m = 0 since all reasonable alignments will have similar, low values of r and any tree found for m = 0 is essentially a tree clustered on r.

Other alignments investigated gave little difference. Fig.9 shows an alternative which is a suitable alternative to B in Fig.6. The results are very similar and illustrate a point. The five aligned cysteines of alternative C might reflect a functional equivalence while the alignment of alternative B might reflect an ancestral equivalence (i.e. true homology). Another possible example occurs in the previously discussed flavodoxins. The gaps at the amino terminal end suggest the most mutationally economical relationship and therefore our best estimate of the homologous residues but Martha Ludwig (personal communication) has shown that, in terms of spatially equivalent residues, the alignment is as if the missing residues were removed from the end. Also, the gap of three residues at positions 101-103 of *Clostridium mp* flavodoxin may be better placed (Ludwig) at positions 98-100 if spatially equivalent residues are aligned. Since this would cost but a few extra nucleotide replacements, it is difficult to decide which is more likely the truly homologous relation. It is therefore a mistake to equate, necessarily, spatial equivalence with homology. After all, the gaps are generally used to relate the various sequences in a genetic sense and since the gaps don't exist in the real sequences, alternatives B and C are identical functionally. By the same token, even though evolution has preserved a near-identical geometry for enzymes of a given class (say, various dehydrogenases), it may be quite erroneous to presume that the spatially equivalent residue positions are occupied by conjugate amino acids[6]. Indeed, if the present-day bacterial ferredoxins derive from the doubling of an ancestral gene as has been suggested by many (Eck & Dayhoff, 1966; Fitch, 1966b; Jukes, 1966; Tanaka et al., 1966; but see Fitch 1973 for some contradictory evidence), the conjugate cysteines in present-day ferredoxins are not used in the same spatial relationships as in their putative progenitor. Adman et al. (1973) have shown by X-ray structural analysis that the four cysteines in the two reactive centers are not the first four and the last four as required by the idea that the spatial equivalence of functional, homologous residues is rigorously preserved in evolution, but rather that one reactive center is composed of the first three and the last cysteine, the remaining cysteines being in the other center, thereby demanding the functional interchange of the fourth and eighth cysteines during evolutionary development. Rossman (1975) has also recognized the potential confusion and wisely chosen to speak of functionally equivalent residues, thereby avoiding the necessary implication of a common ancestral codon entailed in referring to such residues as homologous.

The failure to find an alignment for the five diverse ferredoxins that is much better than any other alternative is no doubt a general hazard for sequences such as these but an alternative possibility in this instance is that all alignments are necessarily wrong, homologously, because the plant N-terminal ferredoxin sequence may have resulted from a frameshift mutation as suggested by Fitch (1970b). His estimated probability that the two sequences would match by chance as well as they did, was not sufficiently low to inspire great confidence that a frameshift was the cause although Reichert et al. (1974) have recently argued that the estimate was much too conservative.

Five rubredoxin sequences have also been examined. Their pairwise distances are shown in Table 6. For low values of m (m_g<1.12, m_n<0.75), *pasteurianum* is most closely related to *elsdenii*, but at larger values it is most closely related to *aerogenes* (see Fig.10). For values of m>4.34, these three are more closely related to the last third of *oleovorans* than to the first third, but for larger values of m they are more closely related to the first third. There is therefore little basis for suggesting which third, if either, more closely reflects the structure and function of the shorter sequences.

[6] see footnote 2.

Table 6. Pairwise distances for 5 bacterial rubredoxins

		1	2	3	4	5
1	*P.aerogenes*	-	38.6	54.9	76.3	65.8
2	*Cl.pasteurianum*	36.2	-	45.4	60.6	60.8
3	*Pe.elsdenii*	43.5	35.8	-	63.5	56.8
4	*Ps.oleovorans*-F	66.9	52.1	50.6	-	69.0
5	*Ps.oleovorans*-L	50.8	47.5	48.2	54.3	-

The values below the principle diagonal are the original distances adjusted for all unknown amino acids (d_u, see text). Those above this diagonal are the original distances adjusted for the number of residues in non-common gaps (d_n). The mutation value for a gap residue, $m_n = 3$. The references for the sequences are: 1, *Peptococcus aerogenes,* Bachmayer et al. (1968a) ; 2, *Clostridium pasteurianum,* McCarthy (1972); 3, *Peptostreptococcus elsdenii,* Bachmayer et al. (1968b); and 4, *Pseudomonas oleovorans;* Benson et al. (1971).

The alignment used was that of Yasunobu & Tanaka (1973b) except that the *aerogenes* gap near residue 30 was interchanged with the following alanine on the grounds that an alanine at an otherwise unvaried threonine is more conservative than an outright deletion and that a gap in a position already containing an aspartate, a lysine and a proline is more likely to be acceptable there than at the otherwise unvaried threonine. The two *Ps.oleovorans* sequences represent the first (-F) and last (-L) thirds, approximately, of a single sequence. These two parts are connected by 62 other amino acids that show no particular homology to these sequences. The best tree, with altered root location, for the d_u data is shown in Fig.10.

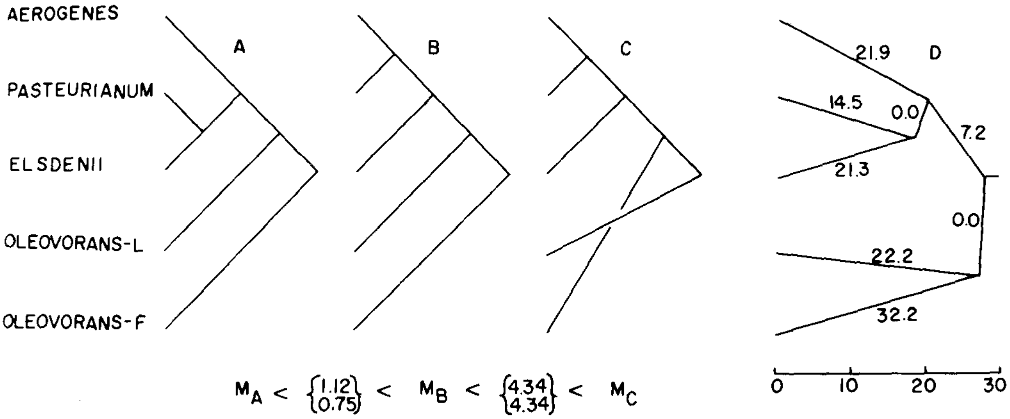
AEROGENES
PASTEURIANUM
ELSDENII
OLEOVORANS-L
OLEOVORANS-F
A
B
C
D
21.9
14.5
0.0
7.2
21.3
0.0
22.2
32.2
0
10
20
30
$M_A < \left\{\begin{matrix}1.12\\0.75\end{matrix}\right\} < M_B < \left\{\begin{matrix}4.34\\4.34\end{matrix}\right\} < M_C$

Fig.10. Bacterial rubredoxin phylogenies. Each tree, A–C, is representative only for the branching order, branch lengths having significance only in tree D. Limiting values of m_g (upper) and m_n (lower) for distances yielding the various topologies is given beneath the trees, A–C. Tree D is the best tree, except for shifting the location of the root (i.e. of the ultimate ancestor), for the d_u data of Table 5. Except for the relocation of the root, it is equivalent in form to tree A

But regardless of the value of *m*, the first and last third of *oleovorans* rubredoxin seem less related to each other than to the other three sequences. This too is interesting because it leads to an impossible phylogeny in which the root lies directly on the path connecting the two *oleovorans'* sequences. But, as parts of the same molecule, they are paralogous in their relationship to each other and the bifurcation between them should represent a non-conjugate, orthologous crossing-over (a partial, internal, gene-duplication event). This is only compatible with the general topology of these sequences if the root is shifted to a point on the link between the ancestor of the two *oleovorans* sequences and that of the other three. This is shown in the right-most tree of Fig.10. Clearly the tree is not unreasonable if the non-conjugate orthologous crossing-over occurred shortly after its separation from the pseudomonad line. Additional rubredoxin sequences might well test that hypothesis.

There are two goals in the development of these procedures. One is to determine whether distances are best calculated on the basis of d_u, d_g, d_n, or to find the conditions under which one is preferable to another. The second is to find those values of m_g and m_n that most often give distances that are likely to yield the correct phylogeny. Those goals are still somewhat distant and will only come closer when sequences with similar problems can be examined from taxa whose phylogenies are more confidently known. In the mean time we can suggest that values for m_g and m_n of 3 and 5 respectively do not seem unreasonable but that even when we know such values better, the sensitivity of the results to the data, the method and the gap values should keep us cautious with respect to all interpretations of those results.

Acknowledgement. This research was supported by the National Science Foundation No. GB 32274X and GB 18739 and the National Institutes of Health Grants GM-16228 and GM-16784.

REFERENCES

Adman, E.T., Sieker, L.C., Jensen, L.H. (1973). J.Biol.Chem. 248, 3987-3996

Aggarwal,S.J., Rao, K.K., Matsubara, H. (1971). J.Biochem.(Tokyo), 69, 601-603

Arnheim, N., Steller, R.(1970). Arch.Biochem.Biophys. 141, 656-661

Azari, P., Tsunoda, J., Glantz, M., Mayhew, S., Yasunoba, K.T. (1970). unpublished

Bachmayer, H., Benson, A.M., Garrard, W.T., Yasunobu, K.T., Whiteley, H.R. (1968a). Biochem. 7, 986-996

Bachmayer, H., Mayhew, S., Peel, J., Yasunobu, K.T. (1968b). J.Biol. Chem. 243, 1022-1030

Benson, A.M., Mower, H.F., Yasunobu, K.T. (1967). Arch.Biochem.Biophys. 121, 563-575

Benson, A.M., Tomoda, K., Chang, J., Matsueda, G., Lode, E.T., Coon, M.J., Yasunobu, K.T. (1971). Biochem.Biophys.Res.Comm. 42, 640-646

Benson, A.M., Yasunobu, K.T. (1969). J.Biol.Chem. 244, 955-963

Dayhoff, M.O. (1972). Atlas of protein sequence and structure, Vol.V, p. D-39. Washington D.C.: National Biomedical Res. Center

Dubourdieu, M., Legall, J., Fox, J.L. (1973). Biochem.Biophys.Res.Comm. 52, 1418-1425

Eck, R.V., Dayhoff, M.O. (1966). Sci. 152, 363-366

Fitch, W.M. (1966a). J.Mol.Biol. 16, 9-16

Fitch, W.M. (1966b). J.Mol.Biol. 16, 17-27

Fitch, W.M. (1969). Biochem.Genet. 3, 99-108

Fitch, W.M. (1970a). Biochem.Genet. 49, 1-14

Fitch, W.M. (1970b). Biochem.Genet. 49, 15-21

Fitch, W.M. (1971). Syst.Zool. 20, 406-416

Fitch, W.M. (1973). Ann.Rev.Genet. 7, 343-381

Fitch, W.M., Margoliash, E. (1967). Sci. 155, 279-284

Fitch, W.M., Margoliash, E. (1970). Evol.Biol. 4, 67-109

Fox, J.L., Smith, S.S., Brown, J.R. (1972). Z.Naturforsch. 27b, 1096-1100

Jukes, T.H. (1966). Molecules and evolution. N.Y.: Columbia Univ. Press 285 pp.

Keresztes-Nagy, S., Perini, F., Margoliash, E. (1969). J.Biol.Chem. 244, 981-995

Matsubara, H., Sasaki, R.M., Chain, R.K. (1967). Proc.Nat.Acad.Sci. 57, 439-445

Matsubara, H., Jukes, T.H., Cantor, C.R. (1968). Brookhaven Symp.Biol. 21, 201-216

Matsubara, H., Sasaki, R.M., Tsuchiya, D.K., Evans, M.C.W. (1970). J. Biol.Chem. 245, 2121-2131

McCarthy, K.T. (1972). Ph.D.Thesis, George Washington University

Moore, G.W., Barnabas, J., Goodman, J. (1973). J.Theoret.Biol. 38, 459-485

Needleman, S., Wunsch, C.D. (1970). J.Mol.Biol. 48, 443-453

Rall, S.C., Bolinger, R.E., Cole, R.D. (1969). Biochem. 8, 2486-2496

Rao, K.K., Matsubara, H. (1970). Biochem.Biophys.Res.Comm. 38, 500-506

Reichert, T.A., Cohen, D.N., Wong, A.K.C. (1974). J.Theoret.Biol. 42, 245-262

Rossman, M.G., Liljas, A., Bränden, C.-I., Banaszak, L.J. (1975). The enzymes, 3rd edition (in press)

Sneath, P.H.A., Sokal, R.R. (1973). Numerical taxonomy. San Francisco: W.H.Freeman and Co. 573 pp.

Sugeno, K., Matsubara, H. (1969). J.Biol.Chem. 244, 2979-2989

Tanaka, M., Nakashima, T., Benson, A.M., Mower, H., Yasunobu, K.T. (1966). Biochem. 5, 1666-1680

Tanaka, M., Haniu, M., Matsueda, G., Yasunobu, K.T., Himes, R.H., Akagi, J.M., Barnes, E.M., Devanathan, T. (1971a). J.Biol.Chem. 246, 3953-3960

Tanaka, M., Haniu, M., Matsueda, G., Yasunobu, K.T., Mayhew, S., Massey, V. (1971b). Biochem. 10, 3041-3046

Tanaka, M., Haniu, M., Yasunobu, K.T., Himes, R., Akagi, J. (1973a). J.Biol.Chem. 248, 5215-5217

Tanaka, M., Haniu, M., Yasunobu, K.T., Mayhew, S.G., Massey, V. (1973b). J.Biol.Chem. 248, 4354-4366

Tanaka, M., Haniu, M., Yasunobu, K.T., Mayhew, S.G. (1974a). J.Biol.Chem. 249, 4393-4396

Tanaka, M., Haniu, M., Yasunobu, K.T., Jones, J.B., Stadtman, T. (1974b). Biochem. 13, 5284-5289

Tanaka, M., Haniu, M., Yasunobu, K.T., Evans, M.C.W., Rao, K.K. (1974c). Biochem. 13, 2953-2959

Travis, J., Newman, D.J., Le Gall, J., Peck, H.D.Jr. (1971). Biochem. Biophys.Res.Comm. 45, 452-458 (1971)

Tsunoda, J.N., Yasunobu, K.T., Whiteley, H.R. (1968). J.Biol.Chem. 243, 6262-6272

Yasunobu, K.T., Tanaka, M. (1973a). In: Iron-sulfur proteins, Vol.II, W.Lovenberg, ed., p. 27-130. N.Y. Acad. Press

Yasunobu, K.T., Tanaka, M. (1973b). Syst.Zool. 22, 570-587

Dr.Walter M.Fitch
Dept. of Physiological Chemistry
University of Wisconsin Medical School
Madison, Wisc. 53706, USA

Note Added in Proof. The statement in the methods section that there is no method for the simultaneous alignment of more than two sequences is no longer true. Sankoff et al. (Sankoff, D., Cedergren, R.J., Lapalme, G., submitted for publication) have a procedure for nucleotide sequences that will in principle handle any number of sequences simultaneously. In practice, limitations on computing time restrict the procedure to only three sequences simultaneously. However, if one is willing to assume a specific phylogenetic relationship among the taxa from which the sequences derive, a recursive operation on the complete set taken 3 taxa at a time converges rapidly to an alignment that must be very close to optimal.