Techniques for the Verification of Minimal Phylogenetic Trees Illustrated with Ten Mammalian Haemoglobin Sequences

David PENNY,* M. D. HENDY† and L. R. FOULDS†

*Department of Botany and Zoology, and †Department of Mathematics, Massey University, Palmerston North, New Zealand

(Received 17 October 1979)

We have recently reported a method to identify the shortest possible phylogenetic tree for a set of protein sequences [Foulds, Hendy & Penny (1979) *J. Mol. Evol.* 13, 127–150; Foulds, Penny & Hendy (1979) *J. Mol. Evol.* 13, 151–166]. The present paper discusses issues that arise during the construction of minimal phylogenetic trees from protein-sequence data. The conversion of the data from amino acid sequences into nucleotide sequences is shown to be advantageous. A new variation of a method for constructing a minimal tree is presented. Our previous methods have involved first constructing a tree and then either proving that it is minimal or transforming it into a minimal tree. The approach presented in the present paper progressively builds up a tree, taxon by taxon. We illustrate this approach by using it to construct a minimal tree for ten mammalian haemoglobin α -chain sequences. Finally we define a measure of the complexity of the data and illustrate a method to derive a directed phylogenetic tree from the minimal tree.

It is well known since the work of Fitch & Margoliash (1967) and Eck & Dayhoff (1966) that amino acid sequences can be used to construct evolutionary trees. It has not been possible previously to decide whether a particular tree represents only a local minimum or whether it is a global minimum. Any conclusions that could be drawn from a particular case have always been subject to the criticism that a shorter tree may be found for the data. Probably this is one of the reasons why these results have not been widely accepted. Hence a global minimal tree is of value for two reasons. Firstly, for comparative purposes it is a mathematically proven benchmark for trees constructed by other methods. Secondly, it provides a mathematical solution to the model of minimizing the number of changes in an evolutionary tree. There is no guarantee of course, that this solution represents what actually happened in evolutionary history.

We have described a new two-part technique (Foulds *et al.*, 1979a,b) that allows the identification of a minimal tree that can link a set of protein sequences. The first part is an efficient tree-building algorithm. The length of the constructed tree is an upper bound on the length of any minimal solution since a minimal tree cannot be longer than any tree connecting the sequences. The

second part independently analyses the data to determine a lower bound on the length of any tree. No tree can exist with a length shorter than the lower bound. The analysis keeps improving the lower bound and searching for shorter trees until the upper and lower bounds are equal. The final tree is then a tree of minimal length.

The present paper presents some improved techniques for the construction and verification of minimal phylogenetic trees from protein-sequence data. It is laid out as follows. The first section presents a method for converting a given data set of amino acid sequences to nucleotide sequences. Some problem cases that arise are dealt with and it is explained why it is desirable to make the conversion. The main section reveals a new strategy for constructing a minimal phylogenetic tree from a set of nucleotide sequences. In previous papers (Foulds *et al.*, 1979a,b; Hendy *et al.*, 1978) we have first built up a tree, taxon by taxon, using a variation of a method described by Prim (1957). We then examined the final tree and attempted to prove it was minimal by partitioning the data into subsets. If this was not successful we tried to transform the tree into a shorter one.

We no longer use this two-stage approach but our strategy is to progressively build a tree, taxon by taxon, and attempt to show minimality at each step.

The advantages of this approach are detailed in the section. The strategy is illustrated by finding a minimal tree for ten mammalian haemoglobin α -chain sequences. The final section of the paper introduces a measure of complexity of the data in constructing a minimal tree from a given data set.

The Conversion to Nucleotide Sequences

We have pointed out in a previous paper (Foulds *et al.*, 1979a) that it is possible to use a criterion of minimizing either the total number of weighted amino acid changes or the number of nucleotide substitutions when constructing a phylogenetic tree based on amino acid sequences. Several workers have converted amino acid sequences into nucleotide sequences before constructing phylogenetic trees (e.g. Fitch & Margoliash, 1967; Moore *et al.*, 1973). In dealing with complex problems we have found it advantageous to convert amino acid sequences into nucleotide sequences. This is because the difference between any pair of nucleotides is of unit weight, which leads to easier partitioning strategies. The method we use to convert the data into nucleotide sequences is now outlined with an example.

For each protein site we list the amino acids that occur at that site. As an example, suppose that only glycine, histidine, leucine, and proline occur in different sequences at a particular site. The triplets that code for these amino acids are given below:

Amino acids		Triplets
Glycine	(Gly)	GGX
Histidine	(His)	CAY
Leucine	(Leu)	CUX or UUR
Proline	(Pro)	CCX

where any of the four nucleotides may occur in the position of an X, A or G may occur in the position of an R, and U or C may occur in the position of a Y (Fitch & Farris, 1974). Each column of the data is then examined independently. Assuming first that CUX codes for leucine, in this case we find that only C and G occur, hence there must be a C–G substitution present in any tree. The second column has four nucleotides, which must require three substitutions. The third column does not require any substitutions since only X and Y occur and therefore either U or C can occur at all positions in the column. Thus any tree will require at least four (1 + 3 + 0) substitutions for this amino acid site, assuming that CUX codes for leucine in this case.

The procedure is repeated with UUR replacing CUX for leucine. However with the choice of UUR it can be seen that any tree will require at least six (2 + 3 + 1) nucleotide substitutions for this amino

acid site. As our criterion is one of minimizing the total number of nucleotide substitutions in the final tree, we assume that CUX codes for leucine in this case.

Arginine and serine are also coded for by any of their six triplets. If more than one amino acid coded for by six triplets occurs at an amino acid site, then all possibilities are considered. The combination corresponding to the smallest number of substitutions is used. Ties are settled by choosing an X in preference to a Y or R, where possible, with remaining ties being settled arbitrarily. If a terminating triplet (UAR or UGA) is present, both possibilities are considered in the same way as above. A terminating triplet is used at the free C-terminus of a protein sequence that is shorter than other sequences in the data set. There could be up to 16 combinations of codes to test if arginine, leucine, serine and a terminating code all occurred at the same amino acid site.

Deletions at the N-terminus are treated differently. For example, two different sequences may have at the N-terminus:

Met-Gly-Arg-His-Lys-Asn-Gly ...
Lys-Asn-Gly ...

The criterion we use for minimizing the tree is to find the smallest number of changes that will connect all the sequences. In the second sequence the position corresponding to that of His in the first sequence is coded for by a deletion symbol (/). This is considered to be one nucleotide change distant from all other codings. Positions to the left are replaced by a null character, which is not compared with any other character. Consequently the two sequences shown are considered to be one change apart for the positions shown.

It is indeed possible that in the course of the evolution of the two sequences there may have been more than one change for the positions shown. Governed by our criterion of minimizing the number of changes, we must select the possibility corresponding to this single change. The same procedure is used for internal deletions. To summarize, the method converts the protein sequences into nucleotide sequences with the minimum total number of changes between the nucleotide sequences.

These nucleotide sequences are used to construct a phylogenetic tree by methods described below. At this stage of our method, at any site a particular amino acid is coded for by the same triplet of nucleotides for all taxa in which it occurs. However it may be possible later to shorten the length of a specific tree by using alternative codes for a particular amino acid for different taxa. An example is given in Fig. 1, where the tree shown is concerned with a single amino acid site and is part of a tree derived by considering complete sequences. Note

Fig. 1. Shortening the length of a phylogenetic tree by using alternative codings for an amino acid

that leucine is present at two points in the tree. The codes shown: CUX at position (1) and UUR at position (2) lead to a tree of length 7. If the same code, either CUX or UUR is used at both positions (1) and (2) the minimal tree is of length 8. If UUR is coded at position (1), CUX at positions (2) the minimal tree is of length 9. Other workers have developed algorithms to produce optimal codings for a given tree (Moore *et al.*, 1973; Fitch & Farris, 1974). After we have constructed a minimal tree for the derived data, changes to the codings are then made whenever the length of this tree can be shortened. This final step cannot be carried out until a tree is constructed.

Therefore there are three steps in the process. First codings are assigned that require the fewest changes, before any tree is constructed. In the second step these codings are used to construct a tree with methods that we have found usually produces minimal trees. The third step involves revising the codings whenever this leads to a decrease in the total length of the tree.

However, it is conceivable that different codings might have resulted in a shorter tree (of different topology) than that finally obtained. This possible problem occurs only when the data are converted from amino acid into nucleotide sequences. The proof of optimality used below is that no shorter tree can be constructed by using the derived data set.

The Π_i strategy

In our experience of tree construction and subsequent proof of minimality, we have found the most difficult phase has been proof of minimality. Hence we have adapted the tree-construction algorithm in order to assist this analysis. We use what are known as 'partitions' of the data set. A partition of the set S of sites of the data set is a collection of subsets of S with the property that every site appears in precisely one subset. Our strategy is to let the tree 'grow', i.e. to link one new taxon at each stage to the tree of the previous stage, and attempt to find a

suitable partitioning of the data set to prove that this new tree is minimal. The major advantage of this approach over our previously described algorithm (Foulds *et al.*, 1979a) is that the partitions 'grow' with the tree. For example, suppose T_i is a tree linking i taxa, which has been proved minimal by the partition Π_i of the data set. Then the necessary partitioning Π_{i+1} to prove T_{i+1} minimal (where T_{i+1} is obtained from T_i by linking one further taxon) can usually be derived simply from Π_i .

This version of our tree-construction strategy is based loosely on the method described by Prim (1957) of constructing minimal spanning trees. We always retain a connected tree T_i at each stage, and the tree T_{i+1} is derived by linking the closest taxon to the tree T_i that is not part of T_i . The major difference from Prim's (1957) algorithm is that we are constructing a minimal phylogenetic spanning tree, which allows for the introduction of Steiner points, i.e. new points at which the tree branches. These new points represent definite sequences that were not present in the original data set (Hendy *et al.*, 1978; Foulds *et al.*, 1979a). In Prim's (1957) algorithm, such an addition is not permitted. Except in the simplest cases, when the two are equal, the Prim (1957) tree will be of greater length than the minimal phylogenetic spanning tree.

A secondary difference is that we do not necessarily choose the shortest link to create our initial tree, T_2 . Instead we select a pair of taxa that are relatively close to each other and to many other species. From experience we have found this to be more successful than the initial step of Prim's (1957) algorithm, which chooses the shortest link. This gives a more uniform growth to T_i and Π_i . We have applied the strategy only to data that have been converted into nucleotide sequences by using the process described above. The tree we produce is a tree of minimal length that spans the derived nucleotide sequences. We have called this development the Π_i strategy.

We illustrate the Π_i strategy with an example, finding an optimal tree for the data set of mammalian haemoglobin α -chain nucleotide sequences listed in Table 1. These sequences were derived by using our conversion algorithm from protein sequences, which come from various sources. We have condensed the sequences by deleting those sites at which there is no variation, and renumbering the site numbers. From a table of differences between taxa, we note the shortest link is MAN-APE, of length 1; however, we soon discover, if we commence the tree building with this link, that MAN-APE and MKY are a relatively isolated part of the tree. Thus we found it more fruitful for the Π_i analysis to begin with the COW-EWE link. (The three-capital-letter abbreviations for the animals are defined in Table 1.)

Table 1. Data after conversion into nucleotides

Invariant sites are omitted. The site number is the amino acid number in the sequences followed by the position within the triplet (in parentheses). For example 4(1) is the first position in the triplet coding for the fourth amino acid. Abbreviations used: A, adenine; C, cytosine; G, guanine; U, uracil; R, A or G; Y, U or C.

Taxon	Abbrev.	Site no.	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4									
Man (<i>Homo sapiens</i>)	MAN	...	4(1)	8(1)	8(2)	10(1)	13(1)	15(2)	17(1)	19(2)	21(2)	22(2)	23(3)	32(2)	34(2)	35(1)	53(2)	55(1)	57(2)	60(1)	64(2)	67(1)	68(1)	68(2)	68(3)	71(1)	73(1)	76(1)	78(1)	78(2)	79(1)	82(1)	82(2)	82(3)	111(1)	113(2)	115(1)	116(3)	124(2)	131(1)	131(2)				
Gorilla (<i>Gorilla gorilla</i>)	APE		C	A	C	G	G	G	C	C	G	R	U	A	C	G	G	A	A	A	A	A	A	A	A	Y	C	G	A	A	A	G	C	R	G	C	U	G	C	R	G	A	G		
Rhesus monkey (<i>Macaca mulatta</i>)	MKY		C	A	C	G	G	G	C	C	G	R	U	A	C	G	G	A	A	A	A	A	A	A	A	U	R	G	G	A	A	A	G	C	R	G	C	U	G	C	R	G	A	G	
Mouse (<i>Mus musculus</i>)	MUS		G	A	G	A	G	G	C	R	U	C	A	C	G	G	A	A	A	A	A	A	A	A	A	Y	C	C	C	G	G	G	G	C	R	U	C	A	G	C	Y	G	A	G	
Dog (<i>Canis familiaris</i>)	DOG		C	A	C	A	A	A	G	C	G	Y	C	A	C	G	C	A	A	A	A	A	A	A	A	C	R	C	C	C	G	G	G	C	R	U	G	A	C	R	G	G	C		
Horse (<i>Equus caballus</i>)	HOS		G	A	C	G	G	G	C	G	R	U	G	C	G	C	A	A	A	A	A	A	A	A	C	U	R	G	C	C	G	G	G	A	Y	G	U	U	A	A	Y	A	G	A	G
Pig (<i>Sus scrofa</i>)	PIG		G	G	C	G	G	C	C	G	R	U	G	A	G	C	C	A	A	A	A	A	A	A	A	R	G	C	A	G	G	G	C	R	U	C	U	A	A	Y	A	A	A	A	
Cow (<i>Bos primigenius</i>)	COW		G	G	G	G	G	C	C	R	U	A	C	G	G	C	A	A	A	A	A	A	A	A	A	R	A	C	C	G	G	G	A	R	U	C	U	U	C	Y	G	A	A	A	
Sheep (<i>Ovis ammon</i>)	EWE		G	A	G	G	G	G	C	C	R	U	A	C	G	G	C	A	A	A	A	A	A	A	A	R	G	C	C	G	G	A	G	A	Y	U	G	U	A	A	Y	A	G	A	A
Rabbit (<i>Oryctolagus cuniculus</i>)	RAB		C	A	C	A	G	A	G	G	R	U	G	A	A	C	A	A	A	A	A	A	A	A	A	R	G	C	C	G	G	A	C	R	A	A	U	U	C	R	G	A	A	A	
Kangaroo (<i>Macropus kanguru</i>)	ROO		G	G	G	A	G	G	C	G	R	C	A	C	A	C	A	A	A	A	A	A	A	A	A	G	C	A	R	A	A	C	G	A	A	R	G	C	U	G	A	R	G	G	C

We have shown (D. Penny, M. D. Hendy & L. R. Foulds, unpublished work) how to prove optimality of a tree with no more than five taxa directly, so T_2 , T_3 (trivially optimal by construction), T_4 , and T_5 are easily proven optimal, and so T_6 is the first case requiring partitioning. The tree T_6 of length 50 connecting EWE, HOS, PIG, MUS, MKY and

The partition is listed with the length of each subset (in nucleotide changes), followed by the sites included in that subset. The first line includes all sites that are in subsets by themselves; that is each subset in the first line has only one site. There are 19 such subsets, of which three (23, 26, 36) require two nucleotide changes each on the tree.

$$\begin{array}{l} \pi_6 = (22) \quad \{1\} \quad \{4\} \quad \{8\} \quad \{9\} \quad \{10\} \quad \{14\} \quad \{16\} \quad \{22\} \quad \{23\} \quad \{25\} \\ \quad \quad \{26\} \quad \{27\} \quad \{29\} \quad \{30\} \quad \{31\} \quad \{36\} \quad \{37\} \quad \{40\} \quad \{42\} \\ (7) \quad \{2, 3, 24, 33\} \\ (4) \quad \{11, 38\} \\ (3) \quad \{15, 43\} \\ (5) \quad \{18, 28, 35\} \\ (4) \quad \{19, 39\} \\ (5) \quad \{20, 34, 41\} \\ \hline 50 \end{array}$$

COW is illustrated in Fig. 2. This tree is minimal, this being proved via the partition Π_6 . To define Π_6 we list each subset of the partition by the site numbers, followed by the length, in brace brackets ($\{ \}$), of the minimal spanning tree needed to span the taxa on that subset of data. We prove minimality provided the sum of these lengths is equal to the length of T_6 .

From our partitioning theorem (Foulds *et al.*, 1979a) we can state that it is not possible to link these data with fewer than 50 nucleotide changes. But a tree of length 50 already exists (Fig. 2), and therefore the tree in Fig. 2 is the shortest possible way of linking the six species.

Now T_7 , of length 54, is created, by adding MAN, as shown in Fig. 3. All duplications of T_7 except

Fig. 2. T_6 , a tree of length 50 that links the six species

No shorter tree can exist, but the analysis has not excluded the possibility of alternative trees of the same total length. Nucleotide changes are expressed, for example as 1CG, on the MKY-S14 link. MKY will have a C at position 1 (Table 1) and S14 a G. The order in which the changes occurred is not specified on the links. Additional evidence is required (see the section on rooted phylogenetic trees) before it is known whether the change was from C to G or vice versa.

those of the MAN-S5 link are justified by the π_6 partition. Hence it is only the further duplications at sites 3, 9, 25 and 26 that need to be justified. The new partition π_7 is derived from π_6 .

$$\begin{array}{l} \pi_7 = (11) \quad \{4\} \quad \{8\} \quad \{10\} \quad \{14\} \quad \{16\} \quad \{22\} \quad \{31\} \quad \{36\} \quad \{37\} \quad \{42\} \\ \quad (3) \quad \{23, 40\} \\ \quad (4) \quad \{1, 26\} \\ \quad (7) \quad \{2, 3, 27, 33\} \\ \quad (5) \quad \{9, 25, 30\} \\ \quad (4) \quad \{11, 38\} \\ \quad (3) \quad \{15, 43\} \\ \quad (5) \quad \{18, 28, 35\} \\ \quad (4) \quad \{19, 39\} \\ \quad (5) \quad \{20, 34, 41\} \\ \quad (3) \quad \{24, 29\} \\ \hline 54 \end{array}$$

Site 36 in the first line requires two nucleotide changes. Thus the seven species cannot be linked with fewer than 54 changes, T_7 (Fig. 3) has 54 changes, and therefore Fig. 3 is the shortest possible tree.

T_8 of length 55 is obtained by adding APE with no further duplications, so $\pi_8 = \pi_7$ with one further site, 12, requiring one change. Thus T_8 is minimal.

Fig. 4 shows the tree T_9 , which is of length 67, obtained when RAB (rabbit) is added. We have not been able to prove that it is minimal by considering partitions that can be generated from those of π_8 . Hence we have added DOG to create T_{10} . The corresponding partitions are:

$$\begin{array}{l} \pi_{10}: (8) \quad \{5\} \quad \{13\} \quad \{14\} \quad \{17\} \quad \{22\} \quad \{31\} \quad \{32\} \\ \quad (14) \quad \{1, 8, 18, 26, 28, 35\} \\ \quad (9) \quad \{2, 3, 27, 33, 39\} \\ \quad (9) \quad \{4, 11, 16, 38\} \\ \quad (14) \quad \{6, 10, 15, 23, 37, 40, 43\} \\ \quad (7) \quad \{9, 12, 25, 30\} \\ \quad (6) \quad \{19, 36\} \\ \quad (7) \quad \{20, 34, 41, 42\} \\ \quad (4) \quad \{24, 29\} \\ \hline 78 \end{array}$$

Site 14 requires two nucleotide changes in Π_{10} . This partition demonstrated that Fig. 5 is the shortest possible way of linking the ten species. As will be observed π_{t+1} is obtained in each case from π_t by combining some subsets and occasionally breaking up a subset to recombine it in differing ways.

A more detailed explanation about how a partition is shown to be minimal is illustrated in Fig. 6, where a subset is taken from the partition for π_{10} .

The subset is on the fifth line of the partition and includes sites 6, 10, 15, 23, 37, 40 and 43. Fig. 6 includes all the different combinations of codes that occur among the ten species being studied [omitting

kangaroo (ROO)]. Species that are only one nucleotide change apart are linked unless a circuit would be formed. Species linked by links of length 1 are called 1-clusters. The method examines all possible ways of linking the 1-clusters together and no trees of length less than 14 are found. The tree must therefore have at least 14 changes at these sites (Foulds *et al.*, 1979a) and this is the number occurring in Fig. 5. We have not been able to prove T_9 minimal by extending the π_8 partition. However, T_9 has been proven minimal (D. Penny, M. D. Hendy & L. R. Foulds, unpublished work) by extensions of the methods given in Hendy *et al.* (1978).

A Measure of Complexity

It has been shown (Foulds *et al.*, 1979a) that a lower bound on the length of the minimal tree for any set of data can be calculated in the following way. At each site in the data where there is variation, construct a minimal spanning tree for the characters that occur. Then the sum of the lengths of these minimal spanning trees is a valid lower bound on the length of the minimal phylogenetic tree.

Fig. 3. Trees T_7 and T_8

A tree of length 54 now includes MAN, and APE can be included with one more link. Both trees are the shortest possible.

Fig. 4. Tree T_9 , of length 66, with RAB included

This is P_0 and is the total number of nucleotide substitutions required considering each site individually.

$$\text{Let } P_0 = \sum_{i=1}^m (N_i - 1) = \left(\sum_{i=1}^m N_i \right) - m$$

Now as we have assumed that the data have been converted into nucleotide sequences, all distinct characters are exactly one nucleotide substitution apart. If there are N_i distinct nucleotides in column i , then the minimal spanning tree has length $(n_i - 1)$.

duplications on the tree nor will help explain duplications at other sites (Fitch, 1977). For π_{10} these are sites 5, 13, 14, 17, 22, 31 and 32, and when these are omitted, $P_0 = 42$ and $L^* = 70$. Thus:

$$C = \frac{70-42}{42} = 0.667$$

To summarize, a measure C has been defined for which an upper bound can be easily found. This measure often gives a good guide as to how difficult it will be to produce a minimal tree. $C = 0$ represents the case where $L^* = P_0$, a tree in which no duplication is necessary, which is the simplest possible case.

The difficulty in verifying that a given final tree is minimal appears to depend more on the complexity of the data (as previously defined) than on the number of species present. For instance, our previous methods, which produced a minimal tree for a data set with 23 species but complexity 0.010 (Foulds *et al.*, 1979b), failed to identify a minimal tree for the present data set, which has ten species and complexity 0.667. However the previous methods are easier to implement than the π_i approach and hence more desirable to use on data of relatively low complexity. Minimal trees have been constructed for a set of ten species derived by using the Π_i approach on a number of different protein sequence data sets (D. Penny, M. D. Hendy & L. R. Foulds, unpublished work).

Rooted Phylogenetic Trees

Most workers have displayed phylogenies with directed links, with one end of each link being earlier in time than the other end. The model that adopts as its optimality criterion the minimization of the length of the tree is concerned solely with trees of undirected links. However should a directed tree (i.e. with directed links) be desired, it can be constructed from an undirected tree provided additional information is available to determine a root (original ancestor) of the tree (Penny, 1976).

One possible method for determining the root of a tree is to introduce a species with known sequence that is thought to have diverged from the group of species under study before any other divergence within the group. This species would identify the root of the tree. The kangaroo, as a marsupial, is a suitable species for our data set as its haemoglobin sequence is available. The shortest way of connecting the kangaroo (ROO) to the existing tree in Fig. 5 is to join it to the HOS—S2 link, giving a phylogeny with directed links shown in Fig. 7. The nucleotide changes on each link are not shown in Fig. 7, but the same changes occur as in Fig. 1 plus 16 on the ROO—S19 line, giving a total length of 94 for the tree in Fig. 7.

Fig. 7. Introduction of a new species to determine the root of the tree

The direction on the links is now specified. The 1CG change (see Fig. 5) would now go from G on S14 to C on monkey whereas the direction of the change was not specified in Figs. 2–5. The tree is of length 94.

Fig. 8. An alternative directed tree, also of length 94

It should be noted that the tree produced by the addition of the ROO—S19 link is not the only minimal tree with the enlarged set of eleven species. The rooted tree shown in Fig. 8 has the same length of 94 and thus Figs. 7 and 8 are equally likely on these data.

Discussion

It should be remembered that the claim to having found the shortest possible tree for a data set is limited in the present case by possible alternative nucleotide sequence codings from the original amino

acid sequences. This will always be a limitation when amino acid sequences are used, but with the marked improvement in DNA sequencing it is likely that nucleic acids will increasingly be used in the future for determining phylogenetic relationships. The use of nucleic acid sequences will simplify the method by eliminating the transformation of the data and will allow a more positive statement that no shorter trees can exist. Another limitation is that it is not usually possible, except with smaller data sets (Hendy *et al.*, 1978), to determine how many trees exist with the minimal length. The method leads to the conclusion in a specific case 'that no shorter tree can link the data', and does not exclude the possibility that other equally short trees exist. An example is shown in Figs. 7 and 8, where two topologies with the same length link the data.

The present paper describes part of a larger study using the same eleven species but five different polypeptides. It is apparent from this work that it is unwise to claim too much biological significance to the one protein used. For this reason any biological conclusions should be avoided, except to mention that primates form a natural group in all cases and that the rodents (as represented by the mouse) are more closely related to primates than to any of the other groups studied.

The present paper therefore reports an improved method for determining minimal trees with complex data sets. It uses the well known conversion of amino acid sequences into nucleotides. The tree-building method is based on the minimal-spanning-tree method of Prim (1957) rather than on that of

Kruskal (1956), which attempts to determine after each species has been added that the tree is minimal before proceeding further. A measure of the complexity of the data is defined, and this can be used to decide how to approach the problem of determining the shortest tree. The conversion of an undirected into a directed (or rooted) tree is illustrated by using a marsupial (kangaroo) sequence.

It is expected that the use of methods such as these will become increasingly common in the next few years as a result of marked improvements in DNA-sequencing techniques.

References

- Eck, R. V. & Dayhoff, M. O. (1966) *Atlas of Protein Sequence and Structure 1966*, National Biomedical Research Foundation, Silver Springs, MD
- Fitch, W. M. (1977) *Am. Nat.* **111**, 223–258
- Fitch, W. M. & Farris, J. S. (1974) *J. Mol. Evol.* **3**, 263
- Fitch, W. M. & Margoliash, E. (1967) *Science* **155**, 279–284
- Foulds, L. R., Hendy, M. D. & Penny, D. (1979a) *J. Mol. Evol.* **13**, 127–150
- Foulds, L. R., Penny, D. & Hendy, M. D. (1979b) *J. Mol. Evol.* **13**, 151–167
- Hendy, M. D., Penny, D. & Foulds, L. R. (1978) *J. Theor. Biol.* **71**, 441–452
- Kruskal, J. R. (1956) *Proc. Am. Math. Soc.* **7**, 48–50
- Moore, G. W., Barnabas, J. & Goodman, M. (1973) *J. Theor. Biol.* **38**, 459–485
- Penny, D. (1976) *J. Mol. Evol.* **8**, 95–116
- Prim, R. C. (1957) *Bell Syst. Tech. J.* **36**, 1389–1401