J. theor. Biol. (1977) **64**, 199–213

Additive Evolutionary Trees†

M. S. WATERMAN ‡

Idaho State University, Pocatello, Idaho 83209, *U.S.A.*

T. F. SMITH

Northern Michigan University, Marquette, Michigan 49855, *U.S.A.*

M. SINGH

State University of New York, Oneonta, New York 13820, *U.S.A.*

AND

W. A. BEYER

Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545, *U.S.A.*

(*Received* 24 *September* 1975)

Metric trees are dendrograms which show the phylogenetic relationships for a set of contemporary species. These dendrograms have numerical values attached to the branches. If the sum of these values on the branches between any two contemporary species is equal to the dissimilarity between these two species, the metric tree is said to be additive and possess an additive dissimilarity matrix. Metric trees and additive matrices are discussed and the uniqueness of the metric tree for an additive dissimilarity matrix is shown. A simple algorithm is given to generate the metric tree for an additive dissimilarity matrix. This algorithm is extended to non-additive dissimilarity matrices through the use of linear programming. Finally, some results for cytochrome *c* sequences are presented.

1. Introduction

A protein is specified by the sequence of amino acids composing the protein. Since 20 kinds of amino acids are used in proteins, a protein is a finite word over an alphabet of 20 letters. Biology seeks to discover the evolutionary relations among the set of proteins. It postulates that proteins have evolved

†Work performed under an NSF Faculty Research Participation Program at the Los Alamos Scientific Laboratory, Theoretical Division, and also supported by the Atomic Energy Commission.

‡Present address: Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545, U.S.A.

in time past, and this evolution has produced a tree (or trees) whose terminal nodes are the extant proteins. It is a task of biology to decide what the mechanisms and probabilities of the various protein evolutionary processes are. It is a mathematical task then, to produce from existing protein sequences and the postulated mechanisms and probabilities of evolutionary processes, a probable tree of the extant proteins. These matters are discussed in the references, but there is no adequate survey. The closest reference to a survey is Dayhoff (1971). These tree construction methods can also be regarded as part of the theory of clusters or of numerical taxonomy. See Jardine & Sibson (1971).

Recently, Moore, Goodman & Barnabas (1973) have outlined a number of the properties of evolutionary trees reconstructed under the additive hypothesis of Cavalli-Sforza & Edwards (1967) from molecular sequence data. Using some ideas from graph theory they define the neighborhood of a given network (tree topology) in terms of nearest neighbor interior vertex interchanges. Using this definition they propose an algorithm for searching a part of the very large space of all possible networks to obtain a network satisfying the additive hypothesis for a given dissimilarity matrix. Several mathematical results are given. However, they fail to prove the convergence of their algorithm, they do not discuss the uniqueness of a given network resulting in an additive dissimilarity matrix, and they do not discuss the conditions which must be satisfied by a dissimilarity matrix to insure it is additive. Some of these points have been discussed by Buneman (1971) and by Dobson (1974).

We shall discuss metric trees and additive dissimilarity matrices and give a uniqueness proof. We prove convergence of a simple algorithm to generate the metric tree for an additive dissimilarity matrix. We make a comparison of the Moore *et al.* (1973) algorithm with our algorithm.

Finally, we discusss the implications of additivity to the evolutionary reconstruction problem and give an extension of our algorithm to non-additive data through the use of linear programming.

While the additive hypothesis surely is an oversimplification of the process of evolution, it is of the utmost biological importance to fully understand the additive hypothesis. Then it will hopefully be possible, as Goodman, Moore, Barnabas & Matsuda (1974), Moore *et al.* (1973) and this paper each attempt, to carry these insights over to a better understanding of evolution.

2. Definitions

We recall the definition of a metric space M as a non-empty set of elements to which a non-negative real number d_{ij} is assigned to every pair i, j in M

and has the following properties:

$$d_{ij} > 0 \qquad \text{for } i \neq j, \tag{1}$$

$$d_{ij} = 0 \qquad \text{for } i = j, \tag{2}$$

$$d_{ij} = d_{ji} \qquad \text{for all } i \text{ and } j, \tag{3}$$

$$d_{ij} \leqslant d_{ik} + d_{kj} \qquad \text{for all } i, j \text{ and } k. \tag{4}$$

A metric tree (unrooted) is a connected graph with no cycles of n terminal vertices with a non-negative real number, called edge length, assigned to each of the $2n-3$ edges. Each vertex has degree 1 or 3. (There are terminology difficulties caused by allowing an edge to have zero length.) The terminal vertices are exactly those of degree 1.

For a set of n elements, a dissimilarity matrix $\mathbf{D}$ is an $n \times n$ matrix of non-negative real numbers $\mathbf{D} = (d_{ij})$ associated with each pair of elements satisfying conditions (1), (2) and (3). This set of n elements is a metric space if and only if the matrix $\mathbf{D}$ satisfies condition (4).

An $n \times n$ dissimilarity matrix $\mathbf{D} = (d_{ij})$ is said to be additive if there exists a metric tree of n terminal vertices such that the sum of edge lengths along the shortest path between terminal vertex i and terminal vertex j equals d_{ij} for all i and j. Thus an additive matrix is a metric space. The necessary and sufficient condition which the dissimilarity matrix must satisfy to be additive has been proven by Buneman (1971) and later by Dobson (1974). This is the four-point condition: For all sets of four elements, there exists some labeling of the elements, say i, j, k and l such that

$$d_{ij} + d_{lk} = d_{jl} + d_{ki} \geqslant d_{il} + d_{jk} \tag{5}$$

If these six distances are interpreted as the edges of a tetrahedron, then the sums of opposite sides form the sides of an isosceles triangle. This is to be compared with an ultrametric in which every triangle is isosceles.

3. Additive Metric Trees

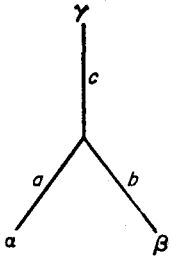

FIG. 1.

Theorem 1

Let **D** be an additive dissimilarity matrix. Then there is one and only one metric tree which yields that dissimilarity matrix.

Proof

Such a metric tree exists by definition and we need only prove its uniqueness. Let α, β, γ be terminal vertices. We wish to find edge lengths a, b, c as shown in Fig. 1. This gives the system of equations

$$d_{\alpha\beta} = a+b, \qquad d_{\alpha\gamma} = a+c, \qquad d_{\beta\gamma} = b+c,$$

with the unique solution:

$$a = (d_{\alpha\beta}+d_{\alpha\gamma}-d_{\beta\gamma})/2,$$
$$b = (d_{\alpha\beta}+d_{\beta\gamma}-d_{\alpha\gamma})/2,$$
$$c = (d_{\alpha\gamma}+d_{\beta\gamma}-d_{\alpha\beta})/2.$$

The triangle inequality assures that a, b and c are non-negative. Therefore there is a unique metric tree for any three terminal vertices.

Now assume, as shown in Fig. 2, there are two metric trees for the terminal vertices α, β, γ, δ.

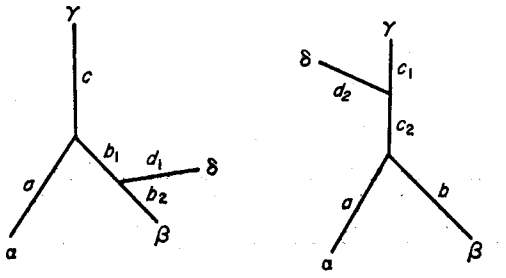

FIG. 2.

One has $b_1+b_2 = b$ and $c_1+c_2 = c$. The assumption of distinct trees implies $b_1+c_2 > 0$. But this implies that the metric tree formed by β, γ, δ is not unique. Thus it is impossible to have two distinct trees for four terminal vertices.

Assume there are $n > 4$ terminal vertices and that there are two distinct tree topologies for these vertices. Each interior vertex of a tree has an associated partition $\{A_1, A_2, A_3\}$ of the terminal vertices and each subset $\{\alpha, \beta, \gamma\}$ of terminal vertices determines a unique interior vertex. In the partition, $A_i = \{\alpha : \alpha$ is a terminal vertex associated with the ith edge connected to the interior vertex$\}$. Let $\{\alpha, \beta, \gamma\}$ determine partition $\{A_1, A_2, A_3\}$

for the first tree and a distinct partition $\{B_1, B_2, B_3\}$ for the second tree. If each of the partitions for every triplet were the same, then the tree topologies would be the same. For each $\varepsilon \in A_i$, there is a j such that $\varepsilon \in B_j$. If $A_i = B_j$ for all ε, the partitions are the same. Thus, for some ε, $\varepsilon \in A_i \cap B_j$ and $A_i \neq B_j$. Let $\delta \in A_j$ and $\delta \notin B_j$. But this is the situation of Fig. 2 and is therefore impossible.

Finally, the edge lengths for the two trees will be shown to be equal. First consider a terminal edge with terminal vertex α and interior vertex v. The vertex v determines the partition $\{\{\alpha\}, A_2, A_3\}$. Choose $\beta \in A_2$ and $\gamma \in A_3$. The metric tree determined by α, β, γ is unique which implies the terminal edge lengths in the two trees are equal.

Next consider an interior edge. Each interior edge partitions the terminal vertices into $\{C_1, C_2, C_3, C_4\}$ where $C_i = \{\alpha : \alpha$ is a terminal vertex associated with the ith edge joining the given interior edge$\}$. Choose $\alpha_i \in C_i$ for $i = 1, 2, 3, 4$. The metric tree determined by $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ is unique which implies the interior edge lengths in the two trees are equal. This completes our proof.

4. The Sequential Algorithm

An algorithm will now be described which allows construction of the unique metric tree for an additive dissimilarity matrix **D**. Pick any two terminal vertices and construct the unique metric tree. Now assume the unique tree has been constructed for a subset C_k of $k \geqslant 2$ terminal vertices. Pick any remaining terminal vertex γ and also α and β from C_k. Find the unique tree with terminal vertices α, β and γ. If the new interior vertex does not coincide with an already existing interior vertex on the path from α to β, it has been added properly to the tree. In case the new vertex does not coincide with an existing vertex on the path from α to β, α or β is replaced by a terminal vertex on the adjoining edge of the tree on which γ must lie, and repeat the above.

Since there are k external vertices, the procedure must be done at most $k-1$ times. (Actually for larger k, one can do much better than that.) Therefore, if the final tree has n terminal vertices, the problem of constructing the tree for three points will be done at most

$$\sum_{k=3}^{n} (k-2) = (n-1)(n-2)/2$$

times. This algorithm finds the unique metric tree since theorem 1 states that each subset of k terminal vertices has a unique tree and this tree must be a subtree of the unique tree for additive **D**. We summarize this result in theorem 2.

Theorem 2

The sequential algorithm described above yields the unique tree for an additive dissimilarity matrix **D**.

We remark here that Moore *et al.* (1973) have an algorithm which seems to converge in the case of an additive **D**. However, they offer no proof of its convergence, a fairly complicated function must be evaluated and minimized for each topology, and they are forced to work, at each step, with networks using all terminal vertices. In their example with ten terminal vertices, it takes ten cycles or an examination of 160 alternative topologies to converge to the unique tree. Using the sequential algorithm we obtain the unique network in eight cycles by a few minutes of hand calculation.

5. Additivity and Molecular Sequence Data

In general, dissimilarity data obtained from molecular sequences such as cytochrome *c* or hemoglobin are non-additive. This appears to be independent of the metric used to generate the dissimilarity measure between sequences. The simplest sequence metric is that defined by Sellers (1974*a*). This metric counts the minimum number of sequence element changes and deletions required to make any two sequences the same. One interpretation of such a number is the total number of mutational events involved in the evolution of any pair of sequences from their most recent common ancestor. Multiple changes at the same site are counted as a single event. Thus, as shown in Fig. 3(a), multiple hits can prevent additivity. The distances for Fig. 3(a) are $d_{12} = d_{14} = d_{23} = d_{34} = 2$ and $d_{13} = d_{24} = 4$. The miltiple hits here are not to be confused with the "multiple hits" identified in real protein sequence data (Fitch & Margoliash, 1967). In those cases, the reference is not to the historical events but to the fact that for some given alignment between three or more molecular sequences, in some position three or more different elements are found. Non-additive data can also result from evolution without multiple hit events as shown in Fig. 3(b). The distances for Fig. 3(b) are $d_{12} = d_{34} = 2$, $d_{23} = 3$, and $d_{13} = d_{14} = d_{24} = 4$. The non-additive condition here is clearly not a result of "misalignment" in measuring the contemporary distances as no deletion or insertion events are involved. The relatively small distance between sequences (2) and (3) can be interpreted as convergent evolution.

The distances in Fig. 3(c) are $d_{12} = d_{13} = d_{34} = 2$, $d_{14} = d_{23} = 3$ and $d_{24} = 4$. These satisfy the four point condition and lead to the additive tree [Fig. 3(d)]. The resulting internal branch lengths are not the original (historical) ones. This is in part a result of allowing non-integer branch lengths.

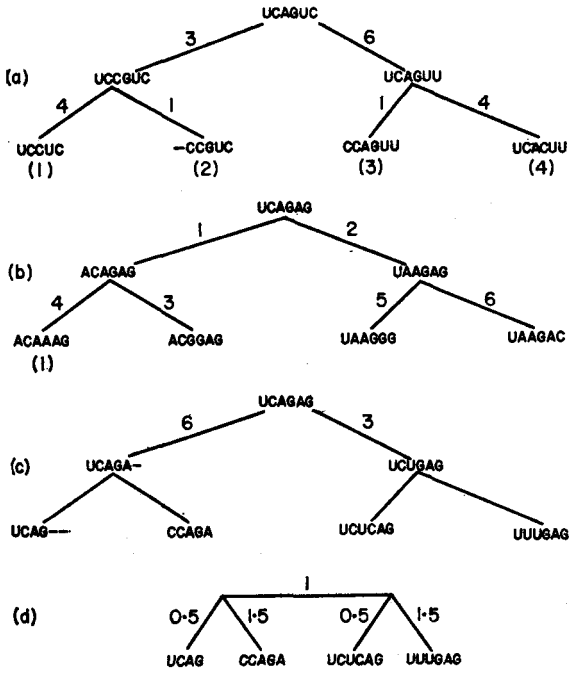

FIG. 3. (a) is an example of multiple hits for which the four point condition is not satisfied. (b) is an example of evolution without multiple hits for which the four point condition is not satisfied. (c) is an example satisfying the four point condition. The numbers along the branches indicate the site changed, counting from left to right. For simplicity there has been only one event along each branch. (d) is the additive tree resulting from the sequential algorithm given in section 4. Here the numbers along the branches are the internal lengths obtained.

Thus for real sequence data in which multiple hits and convergent evolution are highly probable, additivity is not expected and if existent its relationship to the true historical tree is unclear. This is, of course, the case in most protein sequence data. For example, Dayhoff (1973) asserts that among 34 cytochrome *c* sequences studies more than half the sites are thought to have undergone more than one change. This is probably the case in most protein sequence data.

It has been of interest in the past to find the metric tree resulting in a $\mathbf{D}'$ as close to an additive $\mathbf{D}$ as possible. Buneman (1971) has suggested an algorithm for finding such a metric tree. His method results in a $\mathbf{D}' < \mathbf{D}$ which does not seem to be a useful condition, since one reason for non-additivity is the occurrence of multiple hits. Therefore, we conjecture that the condition $\mathbf{D}' > \mathbf{D}$ is more biologically reasonable. Goodman *et al.* (1973)

propose an algorithm apparently satisfying this condition for non-additive data. Their algorithm is a direct extension of their additive data tree reconstruction algorithm. However, as noted earlier, the convergence properties of this approach are unknown.

A justification of these approaches may rest on the idea that for conservative proteins (small rates of change in time) multiple changes have been rare. It is then conjectured that trees which are free of multiple hits are, with a high probability, additive.

6. The Extension of the Sequential Algorithm

An extension of the sequential algorithm which obtains the proper tree for additive dissimilarity matrices is used to study the problems of non-additive dissimilarity matrices. The algorithm for possibly non-additive dissimilarity matrices is as follows. Its justification is that it works in the case of additive dissimilarity matrices. Suppose there are n contemporaries given. The tree is built up as before from an initial metric tree of three terminal vertices by adding successive terminals. Suppose that a tree of $k < n$ terminal vertices has been constructed. The addition of a new edge to all $2k-3$ existing edges is tested. The location of the new edge is determined by finding the minimum value of some linear function of the edge lengths, under some set of linear constraints. This latter procedure is called linear programming [Hadley (1962)] and has been applied to the evolutionary sequence problem by Beyer, Smith, Stein & Ulam (1974). This procedure continues until a tree of n terminal vertices is created.

The weakest meaningful linear constraint appears to be the imposition of the metric properties (1), (2), (3) and (4) on the edge lengths of the tree. This demands that the sum of edge length through the metric tree connecting each pair of terminal vertices be equal to or greater than the corresponding dissimilarity matrix **D**. This last constraint has not been imposed in most earlier studies (Fitch, 1972; Goodman *et al.*, 1974).

The choice of the linear function of edge lengths, properly called the "objective function", is somewhat arbitrary. We have considered three functions of the edge lengths: first the unweighted sum

$$f_1(\{e_i\}) = \sum_{i=1}^{2n-3} e_i$$

of edge lengths. This choice of objective function has been interpreted as minimizing the total evolution. The second objective function used is the sum over the present deviations between the path through the tree connecting

terminal vertices and the corresponding element in the dissimilarity matrix **D**,

$$f_2(\{e_i\}) = \sum_{S_l} \sum_{S_k} \left\{ \sum_{i \in \{S_l, S_k\}} e_i - D_{lk} \right\} \Big/ D_{lk}.$$

Here the sum over $i \in \{S_l, S_k\}$ denotes the sum over the edges in the tree that connect the terminal vertex for sequence S_l to that of S_k. This objective function was first suggested by Fitch & Margoliash (1967). Finally, a third objective function was used,

$$f_3(\{e_i\}) = \sum_{S_l} \sum_{S_k} \left\{ \sum_{i \in \{S_l, S_k\}} e_i - D_{lk} \right\} \Big/ D_{lk}^2$$

differing from the second only in the use of a χ^2-type weighting.

The linear constraints used with all of these objective functions are

$$e_i \geqslant 0 \quad \text{for all } i \tag{6a}$$

and

$$\left[\sum_{i \in \{S_l, S_k\}} e_i \right] \geqslant D_{lk} \tag{6b}$$

for all sequences S_l and S_k, the metric constraint.

It can be shown that any finite solution to the problem occurs at a vertex of the convex polyhedron, whose faces are the boundaries of the half spaces determined by the linear constraints. For example, consider the case of a three by three dissimilarity matrix **D**

$$\mathbf{D} = \begin{pmatrix} 0 & 6 & 2 \\ 6 & 0 & 6 \\ 2 & 6 & 0 \end{pmatrix}.$$

The three triangle inequality constraints

$$e_1 + e_2 \geqslant 6$$
$$e_1 + e_3 \geqslant 2$$
$$e_2 + e_3 \geqslant 6$$

and the non-negativity constraints $e_i \geqslant 0$ confine the solution to at most the four vertices formed in Fig. 4. The co-ordinates of these four vertices are: (6, 0, 6), (1, 5, 1), (2, 6, 0) and (0, 6, 2). The second vertex is the additive solution; i.e. it gives equalities in equation (6) and therefore it minimizes all three of our objective functions, the first one with a value of seven and the second and third with the value of zero. In all four possible solutions at least three of the constraints are satisfied. This is one of the properties of these linear programming solutions: any solution makes a minimum of $2n-3$ constraints into equalities since there are $2n-3$ variables e_i.

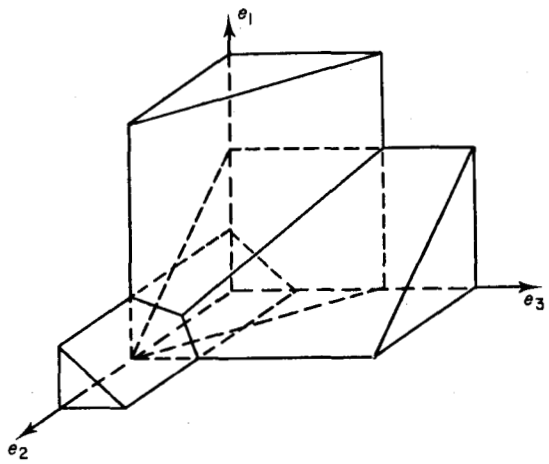

FIG. 4.

The next theorem shows that at least $n-1$ distinct pairs of terminal vertices are additive in the linear programming solution.

Theorem 3

For an $n \times n$ ($n \geqslant 3$) dissimilarity matrix **D**, the assignment of edge lengths e_i by linear programming with constraints equations (6a) and (6b) yields at least $n-1$ of the $\binom{n}{2}$ triangle inequalities as equalities and at most $n-2$ e_i can be equal to zero.

Proof

For $n = 3$, the only topology is shown in Fig. 5.

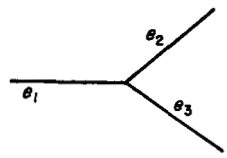

FIG. 5.

At least two of the e_i must be non-zero, for otherwise, say, $0 = e_1 + e_2 \geqslant d_{12}$.

Assume that for n terminal vertices at least $n-1$ e_i must be non-zero. Now consider a tree with $n+1$ terminal vertices. At least one of the associated terminal edge lengths must be non-zero. Otherwise **D** has a non-diagonal

zero element. By the induction hypothesis at least $n-1$ of the e_i in the tree with the non-zero terminal edge length deleted must be non-zero. But $n-1+1 = n = (n+1)-1$ and the induction is completed. Hence for a tree with n terminal vertices at least $n-1$ e_i must be non-zero.

There are a total of $2n-3$ e_i of which at least $n-1$ are non-zero. Thus at most $n-2$ e_i can be zero. This implies that at least $n-1$ of the $\binom{n}{2}$ triangle inequality constraints are equalities.

In summary, it is the metric tree and the associated constraints which determine the vertices of the associated convex polyhedron. Linear programming chooses the vertex of this convex polyhedron which minimizes the objective function.

7. Application to Molecular Sequence Data

Three dissimilarity matrices for 18 contemporary cytochrome sequences were computed using Sellers' (1974*b*) sequence metric. This metric requires a metric on the set of amino acids. The first matrix, $\mathbf{D}_1$, was obtained by the metric which assigns the distance 1 to pairs of unequal amino acids. The matrix $\mathbf{D}_1$ is shown in Table 1 and is the matrix used in the analysis for Tables 2 and 3. The matrix $\mathbf{D}_2$ was obtained using a Fitch & Margoliash (1967) type codon element metric on the set of amino acids and $\mathbf{D}_3$ was obtained using Sneath's (1966) chemical metric. These three D's, the extended algorithm for non-additive dissimilarity matrices given above, and the three objective functions form the basis of our computer analysis. The results are summarized as follows.

First, the final metric tree depends on the order in which the terminal vertices are added. Secondly, the metric trees depend very little on which objective function or which dissimilarity matrix is used. No order of addition for any objective function resulted in a topology in complete agreement with the more accepted evolutionary trees (see Fig. 3 in Beyer *et al.*, 1974). The addition of the rattlesnake later in the algorithm gave results more in agreement with accepted evolutionary trees. The problem of the position of the rattlesnake in an evolutionary tree was noted by Gibbs & MacIntyre (1970) and is believed related to the closeness of the rattlesnake and human cytochrome *c*'s. Table 2 shows the dependence on the order of terminal vertex addition. Following Goodman *et al.* (1974), all trees resulting from the interchange of a single pair of nearest neighbor internal vertices were studied. In all cases in Table 2, the tree found by the extension of the sequential algorithm was a local objective function minimum among these nearest

TABLE 1

Distances among 18 organisms. Distances are obtained from Seller's metric using $\delta(a, b) = 1$ *for amino acids* $a \neq b$

(1) Human	0																
(2) Rhesus Monkey	2	0															
(3) Horse	15	15	0														
(4) Donkey	14	14	1	0													
(5) Pig	11	11	5	4	0												
(6) Dog	12	12	8	7	3	0											
(7) Gray Whale	12	12	8	7	3	4	0										
(8) Rabbit	11	11	9	8	5	6	2	0									
(9) Kangaroo	11	13	9	10	6	7	7	7	0								
(10) Chicken	17	17	14	13	11	12	10	10	14	0							
(11) Penguin	17	17	15	14	12	12	10	10	12	2	0						
(12) Pekin Duck	16	16	14	13	11	10	9	9	13	3	3	0					
(13) Pigeon	15	15	14	13	11	10	9	8	13	5	5	4	0				
(14) Rattlesnake	19	21	30	29	28	28	25	24	28	26	26	23	24	0			
(15) Bullfrog	24	24	20	19	15	16	16	16	18	16	17	16	17	32	0		
(16) Tuna Fish	30	32	27	26	25	28	25	25	26	25	26	25	26	36	21	0	
(17) Dogfish	31	31	20	19	20	23	22	24	24	24	25	24	25	35	27	26	0
(18) Lamprey	24	26	19	18	17	20	20	22	23	24	25	24	24	35	29	26	21	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

TABLE 2

Examples of objective function dependence on order of terminal vertex addition in the extended sequential algorithm. A set of 18 cytochrome c protein sequences were used in this example. Order 1 is: 1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18. Order 2 is the single link cluster order (Jardine & Sibson, 1971): 1-2-3-4-6-7-8-5-9-12-13-10-11-14-18-15-16-17. The third order is obtained from an average linkage cluster analysis (Jardine & Sibson, 1971): 1-2-13-11-10-7-8-6-5-4-3-9-12-14-15-16-17-18, and the fourth order differs only in the point of addition of the rattlesnake: 1-2-13-11-10-7-8-6-5-4-3-9-12-15-16-17-18-14. The distance from the accepted tree is the number of nearest neighbor vertex interchanges required to convert to the accepted evolutionary tree (Beyer et al., 1974)

Order of Addition	f_2 ($\{e_i\}$) (objective function)	Distance from accepted tree
1	15·75	2
2	16·52	4
3	15·61	5
4	14·92	8
Accepted tree	16·13	0

neighbors. However, in other computer runs on a larger number of cytochrome *c* sequences exceptions were found to this local minimum property. Since the extension of the sequential algorithm does not always converge to a global minimum, subtrees of optimal trees are not necessarily optimal.

To circumvent the metric tree dependence on the order of terminal vertex addition, the order of addition was made to depend on the objective function value. This was done by testing the addition of all remaining sequences associated with terminal vertices and adding the one giving the lowest value of the objective function. This was continued until all terminal vertices were added. This approach results in a tree composed of optimal subtrees. Table 3 compares this method with the above. The trees are again nearest neighbor vertex minimum.

In conclusion, we find our metric tree construction algorithms do not necessarily give a global minimum. Therefore, since it is not practical to search all trees for, say, 18 terminal vertices, we recommend searching biologically reasonable trees. As concluded by Buneman (1971) in his study of the properties of additive metric trees, the ability to reconstruct evolution using the sequence metric approach may be limited. It has been noted

TABLE 3

An example comparison of the extended sequential algorithm, orders A through F, with G which is the order obtained by minimizing f_3 ($\{e_1\}$) *for each possible additional terminal vertex. Note: three topologies have lower* f_3*'s including A, the one resulting in the more biological topology. The order numbers refer to the cytochrome* c *listed in Table 1 plus the one for the silkworm (19). The dissimilarity matrix* $\mathbf{D}_1$ *used is given in 3(b)*

	Order of addition	f_3 ($\{e_i\}$) (objective function)
(a)		
A	3– 9– 1–11–15–16–17–18–19	2·459
B	18–17–16–19– 9– 3– 1–11–15	2·459
C	18–17–16–15–11– 3– 9– 1–19	2·547
D	1– 3– 9–15–11–19–16–17–18	2·603
E	19–18–17–16–15–11– 1– 9– 3	3·002
F	19– 9–18–15–11– 3– 1–16–17	3·049
G	1– 3– 9–15–19–16–11–17–18	2·603

(b)

3	0								
9	9	0							
1	15	11	0						
11	15	13	19	0					
15	20	18	24	18	0				
16	28	26	32	26	30	0			
17	27	25	31	25	29	26	0		
18	27	25	31	25	29	26	21	0	
19	36	34	40	34	38	44	43	43	0
	3	9	1	11	15	16	17	18	19

(Dayhoff, 1971) that the $n \times n$ dissimilarity matrix contains considerably less information than the original sequences. This has led some investigators to prefer an ancestral sequence reconstruction approach.

REFERENCES

BEYER, W. A., SMITH, T. F., STEIN, M. L. & ULAM, S. M. (1974). *Math. Biosciences* **19**, 9.

BUNEMAN, P. (1971). *Mathematics in the Archeological and Historial Sciences*, (F. R. Hobson, D. G. Kendall & P. Tautu, eds) p. 387. Edinburgh: University Press.

CAVALLI-SFORZA, L. L. & EDWARDS, A. W. F. (1967). *Am. J. Human Genet.* **19**, 233.

DAYHOFF, M. O. (1971). *Atlas of Protein Sequence and Structure.* Silver Spring: National Biomedical Research Foundation.

DOBSON, A. J. (1974). *J. appl. Prob.* **11**, 32.

FITCH, W. M. (1972). *J. molec. Evolution.* **1**, 185.

FITCH, E. M. & MARGOLIASH, E. (1967). *Science, N.Y.* **155**, 279.
GIBBS, A. J. & MACINTYRE, G. A. (1970). *Eur. J. Biochem.* **16**, 1.
GOODMAN, M., MOORE, G. W., BARNABAS, J. & MATSUDA, G. (1974). *J. molec. Evolution.* **3**, 1.
HADLEY, G. (1962). *Linear Programming.* Reading Mass.: Addison-Wesley.
JARDINE, N. & SIBSON, R. (1971). *Mathematical Taxonomy.* New York: John R. Wiley and Sons.
MOORE, G. W., GOODMAN, M. & BARNABAS, J. (1973). *J. theor. Biol.* **38**, 423.
SELLERS, P. H. (1974*a*). *J. Combinatorial Theory* (A), **16**, 253.
SELLERS, P. H. (1974*b*). *Siam. J. appl. Math.* **26**, 787.
SNEATH, P. H. A. (1966). *J. theor. Biol.* **12**, 157.