J. theor. Biol. (1978) **73**, 789–800

On the Similarity of Dendrograms

M. S. WATERMAN

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

T. F. SMITH

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

(*Received* 21 *November* 1977, *and in revised form* 4 *April* 1978)

A metric on binary trees is defined to give the similarity of two dendrograms. One of the major desirable properties of the proposed tree similarity measure is to clarify the decision ordering nature of biological trees. This metric is applied to evolutionary tree reconstructions and comparative embryogenesis. The mathematical properties of this metric are discussed, and an algorithm is proposed to compute the metric.

"....our essential task lies in the comparison of related forms rather than in the precise definition of each; and the deformation of a complicated figure may be a phenomenon easy of comprehension, though the figure itself have to be left unanalysed...."

D'arcy Thompson 1917

1. Introduction

A number of areas of current biological research result in, and use binary trees (dendrograms). The most evident area is taxonomy. Here various hierarchical cluster methods (Hartigan, 1975; Jardine & Sibson, 1971; Johnson, 1967) are used to construct taxonomic trees. Over the last ten years, rather successful attempts have been made in reconstructing evolutionary trees from molecular sequence data. (See Dobson, 1975; Jardine & Sibson, 1971; Moore, Goodman & Barnabas, 1973; Sneath & Sokal, 1973; Waterman, Smith, Singh, & Beyer, 1977). One of the remaining problems this area is that the different cluster methods result in different (Hartigan, 1975; Jardine & Sibson, 1971) dendrographic structures. Worse yet, the same cluster algorithm can result in different trees with differing initial ordering (Waterman *et al.*, 1977). Therefore, in numerical taxonomy a method of measuring the degree of similarity between dendrograms is of some importance. This is true both for comparative taxonomic studies and

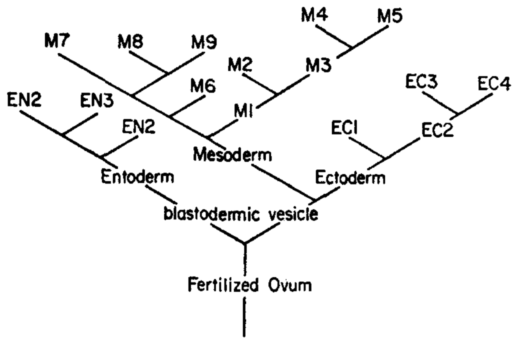

FIG. 1. A dendrographic representation of some of the major progressive tissue differentiations arising from the three primary germ layers in the vertibrate embryo. Tissues obtained from Ectoderm are: EC1 = Pharynegeal pouches, EC2 = Liver, EC3 = Pancreas. Tissues obtained from the Mesoderm are: M1 = lateral mesoderm, M2 = somatic mesoderm, M3 = splanchnic mesoderm, M4 = heart, M5 = blood corpuscles, M6 = skull, M7 = axial skeleton, M8 = skeleton muscles, M9 = skin connective tissue. Tissues obtained from the Entoderm are: EN1 = pharyngeal pouches, EN2 = liver, EN3 = pancreas.

for the continuing development of the dendrographic construction techniques (Dobson, 1975; Farris, 1973; Jardine & Sibson, 1971).

A second area of biological research involving dendrograms is in morphogenesis and/or cell differentiation studies (Reverberi, 1971). As in the dendrogram representation of evolutionary history, the development of many tissue systems may be represented by a tree (see Fig. 1). This representation does not seem to have been exploited previously. These trees arise basically as decision trees, which have been studied extensively by computer scientists (Knuth, 1969; Prather, 1976; Korfhage, 1974). One of the desirable properties of a tree similarity measure would be to clarify the decision order nature of biological trees. With such a measure some aspects of comparative embryogenesis can be put on a more quantitative basis.

The problem of the similarity of dendrograms thus arises in a number of areas. There are two independent properties of binary trees of interest. First is the dendrographic structure (branching topology) and second are the branch lengths. The branch lengths are normally directly proportional to some parameter, such as time, number of mutations or number of cell divisions. The branching topology contains the fundamental information about the order in which the "decisions" were made. It is this second property which we wish to consider in connection with similarities between binary trees. The degree of similarity, in general, is a problem of graph theory (Dobson, 1975). While computer science frequently deals with tree structures (Knuth, 1969; Prather, 1976; Korfhage, 1974), this particular problem does not seem to have been considered in that context.

The subject of this paper, then, is to find a metric ρ on the set, $\mathscr{T}_n$, of unrooted binary trees over the same set of n terminal vertices. Given two trees T and S in the set we require

$$0 \leqslant \rho(T, S) \leqslant 1$$

in addition to the usual requirements that $(\rho, \mathscr{T}_n)$ be a metric space. This requirement allows comparison of two clustering schemes applied to sets with different numbers of terminal vertices.

This topic has been previously surveyed by Dobson (1975) who considered various approaches. The metric proposed in this paper is new and consequently not discussed by Dobson. The new metric properties mentioned below may make it more appealing than these previously proposed. However, the present work lacks the value of the normalizing constant and does not have an efficient algorithm for computation. Hopefully, these problems will be solved in later work. However, an algorithm that can be computed is given, and it is conjectured that it coincides with the metric mentioned above.

2. Binary Trees

We define an *(unrooted)*† *binary tree* or *dendrogram* to be a connected graph with no cycles where each vertex has degree one or three and where the terminal vertices (those of degree one) are labeled. If a binary tree has n terminal vertices, then it has $n-2$ internal vertices, n branches connecting terminal to internal vertices, and $n-3$ branches connecting internal vertices.

In work on constructing evolutionary trees from molecular data sets, Moore, Goodman & Barnabas (1973) introduce the concept of "nearest neighbor 1-step changes." This idea, which we refer to as "nearest neighbor interchange," was used to generate the neighborhood of a given tree in order to search the very large set of possible trees for the tree giving the best fit to a molecular data set. For a discussion of these and related concepts, see Waterman, Smith, Singh & Beyer (1977).

To give a careful definition of nearest neighbor interchange, it is useful to consider some natural partitions of the terminal vertices of the binary tree. Each interior vertex induces a natural partition of the terminal vertices into three sets each of which is composed of those terminal vertices in each connected component after the removal of the interior vertex. More important for the present discussion is the partition associated with each interior branch. Removal of the interior branch, but not the two interior vertices

†For application to rooted trees, one need only to identify a given terminal vertex as the "root".

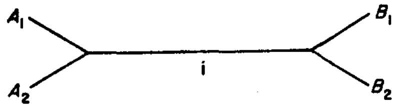

FIG. 2. Separation $\nu = \{\{A_1, A_2\}, \{B_1, B_2\}\}$ associated with interior branch **i**.

associated with the branch, yields the *partition* $\pi = \{A, B\}$. Removal of the vertices yields $\nu = \{\{A_1, A_2\}, \{B_1, B_2\}\}$, where $A_1 \cap A_2 = \phi$, $A_1 \cup A_2 = A$, $B_1 \cap B_2 = \phi$ and $B_1 \cup B_2 = B$. The *separation* ν *of the partition* π is indicated in Fig. 2. where the interior branch is labeled **i**.

The set of partitions for all interior branches, $\mathscr{P} = \{\pi_1, \ldots, \pi_{n-3}\}$, is equivalent to the binary tree generating the partitions. This fact is given in the next theorem.

Theorem 1

If $\mathscr{P}$ is the unique set of partitions for a given binary tree, then $\mathscr{P}$ can be used as a representation of the binary tree.

Proof

Let $\mathscr{P} = \{\pi_1, \ldots, \pi_{n-3}\}$ be the set of partitions for a tree with terminal vertices labeled $1, 2, \ldots, n (n>4)$. Some $\pi_i = \{A_i, B_i\}$ is such that A_i (say) has exactly two elements, $A_i = \{j_1, j_2\}$. Since we assumed $\mathscr{P}$ was constructed from a binary tree, j_1 and j_2 occur together in the sets of all the partitions. Thus the problem can be relabeled to make a new problem with $n-1$ terminal vertices. Continue the process $n-4$ steps where the tree of four terminal vertices corresponds to a unique partition. Of course, the relabeling must be reversed to recover the tree giving rise to $\mathscr{P}$.

Given a separation $\nu_0 = \{\{A_1, A_2\}, \{B_1, B_2\}\}$ associated with an interior branch, a tree is said to result from a *nearest neighbor interchange* if it has separation $\nu_1 = \{\{A_1, B_1\}, \{A_2, B_2\}\}$ or $\nu_2 = \{\{A_1, B_2\}, \{B_1, A_2\}\}$ where all other separations remain the same. These separations are shown in Fig. 3.

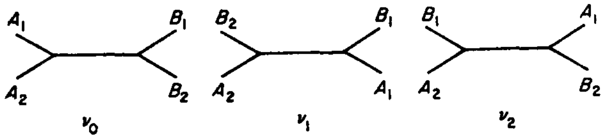

FIG. 3. The three nearest neighbors ν_0, ν_1, ν_2.

Thus a nearest neighbor interchange results in one of two possible trees. There are two equivalent ways of viewing the generation of ν_1 or ν_2 from ν_0.

First we can consider moving the branch associated with A_i (say) onto the branch associated with either B_1 or B_2. Equivalently, these separations can be generated by interchanging A_i and $\mathbf{B}_j$

3. Nearest Neighbor Interchange Metric

The metric we propose below essentially counts the minimum number of nearest neighbor interchanges required to change one tree to another. It should be noted that interchanges deep in the tree are given equal weight with, say, an interchange involving a terminal vertex. In the context of evolutionary trees, this says that misclassification or changed decision order deep in the tree was, at the time that specification or decision occurred, exactly as important as misclassification at a more recent level.

Now, given T and S in $\mathscr{T}_n$, the set of binary trees with the same set $\{1, 2, \ldots, n\}$ of terminal vertices define

$$\{T \to S\} = \{\tau_l\tau_{l-1}\ldots\tau_1 : \tau_l(\tau_{l-1}(\ldots\tau_1(T)\ldots)) = S,$$

$\tau_i : \mathscr{T}_n \to \mathscr{T}_n$, and τ_i is the identity or results in a nearest neighbor interchange, $1 \leqslant i \leqslant l\}$.

If the identity transformation belongs to $\{T \to S\}$, let $l = 0$. Also, it is shown below that $\{T \to S\} \neq \phi$ if $T \neq S$.

Theorem 2

Let T and $S \in \mathscr{T}_n$. Define

$$M(n) = \max_{Q,P \in \mathscr{T}_n} \min\,[l : \tau_l\ldots\tau_1 \in \{Q \to P\}]$$

and

$$\rho(T, S) = \frac{\min\,[l : \tau_l\ldots\tau_1 \in \{T \to S\}]}{M(n)}.$$

Then

(i) $M(n) < \infty$ for $1 \leqslant n < \infty$.

(ii) $0 \leqslant \rho\,(T, S) \leqslant 1$ for $T, S \in \mathscr{T}_n$.

(iii) $(\rho, \mathscr{T}_n)$ is a metric space.

Proof

(i) To show $M(n)$ finite, consider any two trees T and S in $\mathscr{T}_n$ and consider $\{l : \tau_l \ldots \tau_1 \in \{T \to S\}\}$. We will exhibit a member of $\{T \to S\}$ with $l < \infty$ so that $M(n) < \infty$ follows.

Now let $\alpha, \beta \in \{1, 2, \ldots, n\}$ be such that there is only one intervening node on the shortest path between them in S. If this is not the situation in T, move α to β. This can clearly be done in less than $n-2$ nearest neighbor interchanges.

Next, in both trees, replace α, β and the shortest path between them by a new terminal vertex α'. Repeat the above process until there are only three terminal vertices remaining. Thus, we have shown that

$$M(n) \leqslant \sum_{i=3}^{n-1} (n-i) = (n-3)(n-2)/2 < \infty.$$

(ii) We have $\rho(T, S) \leqslant 1$ by definition of $M(n)$.

(iii) Assume $T = S$. Then the identity transformation belongs to $\{T \rightarrow S\}$ and $\rho(T, S) = 0$. Also, if $\rho(T, S) = 0$, then since $M(n) < \infty$, the identity transformation belongs to $\{T \rightarrow S\}$ or $T = S$.

Clearly $\rho(T, S) \geq 0$.

Since a transformation corresponding to a nearest neighbor interchange has an inverse which is a nearest neighbor interchange, $\rho(T, S) = \rho(S, T)$.

Finally, we establish the triangle inequality. Suppose

$$\tau_{l_1} \ldots \tau_1 \in \{T \rightarrow S\}$$

and

$$\tau_{l_2+l_1} \ldots \tau_{1+l_1} \in \{S \rightarrow R\},$$

where

$$l_1 = \min \{l \colon \tau_l \ldots \tau_1 \in \{T \rightarrow S\}\}$$

and

$$l_2 = \min \{l \colon \tau_l \ldots \tau_1 \in \{S \rightarrow R\}\}.$$

Then

$$\tau_{l_2+l_1} \ldots \tau_{1+l_1} \tau_{l_1} \ldots \tau_1 \in \{T \rightarrow R\}$$

so that

$$\rho(T, R) \leq \frac{l_1 + l_2}{M(n)} = \rho(T, S) + \rho(S, R).$$

This completes the proof.

An important relationship of the ultrametric inequality with binary trees is that any additive tree satisfies that inequality (Dobson, 1975; Johnson, 1967). It is natural, then, to ask whether ρ satisfies this inequality:

$$\rho(T_1, T_2) \leqslant \rho(T_1, T_3) = \rho(T_2, T_3)$$

for all triplets of trees with some labeling. This is equivalent to

$$M(n)\rho(T_1, T_2) \leqslant M(n)\rho(T_1, T_3) = M(n)\rho(T_2, T_3).$$

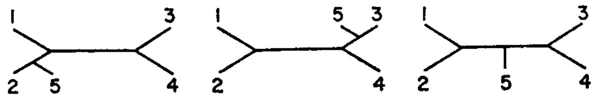

FIG. 4. Example ($n = 5$) where the ultrametric inequality fails.

The inequality is satisfied for $n = 4$ but, for $n = 5$, a counter example is given in Fig. 4, where

$$M(n)\rho(T_1, T_2) = 2 \quad \text{but} \quad M(n)\rho(T_1, T_3) = M(n)\rho(T_2, T_3) = 1.$$

4. Closest Partitions Metric

In this section, we consider an algorithm based on the partitions induced by the interior branches of a binary tree. Essentially, the algorithm searches for the minimum number of nearest neighbor interchanges (*nni*) to achieve a partition in T that agrees with a partition in S. (The problem is to find a minimal member of $\{T \to S\}$.) Hence the algorithm is called the closest partition (CP) algorithm.

As soon as the first identical partition is achieved, a descendent partition from T, π_T, and a partition in S, π_S, agree. That is, $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ are equal for each partition. Therefore, the algorithm branches into two problems with the interior vertex partitions $\{A, B_1, B_2\}$ and $\{A_1, A_2, B\}$, where each A and B is considered a terminal vertex. This procedure is shown in Fig. 5. Note that each new problem has no more than $n-1$ terminal vertices and that convergence of the algorithm is then clear. Since each new tree is essentially a non-overlapping part of the problem, it should be emphasized that the algorithm branches in a notational sense only.

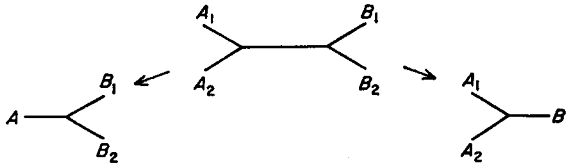

FIG. 5. The original problem with the two smaller descendent problems.

The following three theorems are presented to explore the relationship between the algorithm and the nearest neighbor interchange metric. Although we wished to show the algorithm finds the minimal path, this has not been completely established.

First, we give a general result which suggests the CP algorithm.

Theorem 3

If an optimal path in $\{T \to S\}$ has a partition π_i associated with branch $\mathbf{i}$ identical with a partition in S, then that branch $\mathbf{i}$ is crossed at most once by any member of $\{1, 2, \ldots, n\}$.

Proof

Let $v = \{\{A_1, A_2\}, \{B_1, B_2\}\}$ and suppose $b \in B = B_1 \cup B_2$. To move b more than once across $\mathbf{i}$ will take at least one more move than leaving b on the appropriate side of $\pi = \{A_1 \cup A_2, B_1 \cup B_2\}$.

The next result handles the situation with existing identical partitions.

Theorem 4

If π_T is a partition in T and π_S is a partition in S and $\pi_T = \pi_S$, then π_T will not change in any optimal path in $\{T \rightarrow S\}$.

Proof

Any *nni* across the associated branch would require the elements to be returned across the same branch. This violates Theorem 3.

Next, we consider the case where one *nni* results in an identical partition.

Theorem 5

If there are no identical partitions between T and S and some nearest neighbor interchange in T results in an identical partition with S, then that nearest neighbor interchange is on some optimal path in $\{T \rightarrow S\}$.

Proof

Since the partition in question must eventually be achieved, the following question must be answered: Could we do *strictly* better by making other moves before achieving this partition?

To answer this question, let

$$\pi_T = \{\{A_1, A_2\}, \{B_1, B_2\}\} \rightarrow \pi_{T_2} = \{\{B_2, A_2\}, \{B_1, A_1\}\}$$

and

$$\pi_S = \{\{C_1, C_2\}, \{D_1, D_2\}\},$$

where it is assumed that $B_2 \cup A_2 = C_1 \cup C_2$. Several possible *nni*'s can be made before $T \rightarrow T_2$ and they will be considered in a case by case fashion:

i) Interchanges within A_1, A_2, B_1, or B_2 need not be made until after $T \rightarrow T_2$ since they can then be made at no change in the total number of interchanges.

ii) Interchanges between A_1 and B_1 or between A_2 and B_2 can, after $T \rightarrow T_2$, be made with fewer interchanges.

iii) Interchanges between A_1 and B_1, between A_1 and A_2, between A_2 and B_1, and between B_1 and B_2 all take more moves than if not made. This holds since they are now separated and they need to be separated to achieve $\{T \rightarrow S\}$.

Results concerning a larger number of *nni*'s to achieve the first identical partition with S have not been obtained. However, we have constructed our CP algorithm based on these partial results.

We now summarize the closest partition algorithm:

CP 1. If an existing partition on T agrees with a partition on S, make two subproblems to continue with. (This procedure is shown in Fig. 5). When any problem is reduced to three terminal vertices that problem is finished.

CP 2. Make k-step nearest neighbor interchanges where k is the minimum number required to produce an identical partition between T and S. All such trees produced will be returned to CP 1.

CP 3. $\rho_P(T, S) \cdot M_P(n)$ then equals the sum of k over all cycles through step CP 2.

Since we have not proven that the closest partition algorithm produces the minimum number of steps, define above $M_P(n)$ and $\rho_P(T, S)$ to be quantities for the closest partition algorithm which correspond to $M(n)$ and $\rho(T, S)$. Then we conjecture that

$$\rho_p(T, S) = \rho(T, S).$$

5. Interior Vertex Algorithm

Prior to our work, two methods have been proposed for measuring similarity of trees based on the number of interior vertices on the path between each pair of exterior vertices. For each tree a sequence of $\binom{n}{2}$ integers is obtained. Phipps (1971) computes the correlation coefficient of the two sequences. It has been conjectured [Dobson (1975)] that $-0{\cdot}5$ is the smallest possible value but a proof of this seems to be unknown. Williams and Clifford (1971) calculate the sum of absolute values of differences between members of the sequences. They divide this sum by $\binom{n}{2}$ to give a measure of similarity. However, the quantity is not bounded above by one. It would be interesting to find the maximum value obtained by the Williams & Clifford scheme.

Our objection to both methods described above is that one classification mistake (a nearest neighbor interchange) can make the two trees seem quite dissimilar. However, there is an interesting algorithm which chooses the nearest neighbor interchange reducing the Williams and Clifford sum the most.

First, let $d_T(i, j)$ be the number of interior vertices on the path between exterior vertices i and j. An interesting formula for d_T is

$$d_T(i, j) = n-2-\sum_{v=1}^{n-3} [I_{A_v}(i)\cdot I_{A_v}(j)+I_{B_v}(i)\cdot I_{B_v}(j)],$$

where $\{A_v, B_v\}$ is the partition associated with the v-th interior branch and I is the usual indicator function.

The interior vertex algorithm chooses the nearest neighbor interchange which minimizes

$$\sum_{\substack{\text{pairs}\\(i,j)}} |d_{T'}(i, j)-d_S(i, j)|,$$

where T' is obtained from T by a nearest neighbor interchange. Unfortunately, it has not even been proven that this algorithm converges.

6. Conclusion

The *nni* metric has been applied to tree reconstruction techniques used on molecular sequence data by Waterman *et al.* (1977). In Figure 6, five evolutionary trees obtained from molecular sequence data are given. These represent the results of different clustering techniques and/or different protein sequence data sources. The *nni* metrics recorded in Table 1 support two conclusions. First, the different clustering or tree building techniques employed result in minimal differences, on the order of those arising from using different sequence data sources. The more apparent conclusion, on study of Fig 6, is that the data do not allowus to determine the consanguinity of the Lagomorpha (rabbits) and carnivora (dogs) with the rest of the phylum chordata (vertibrate).

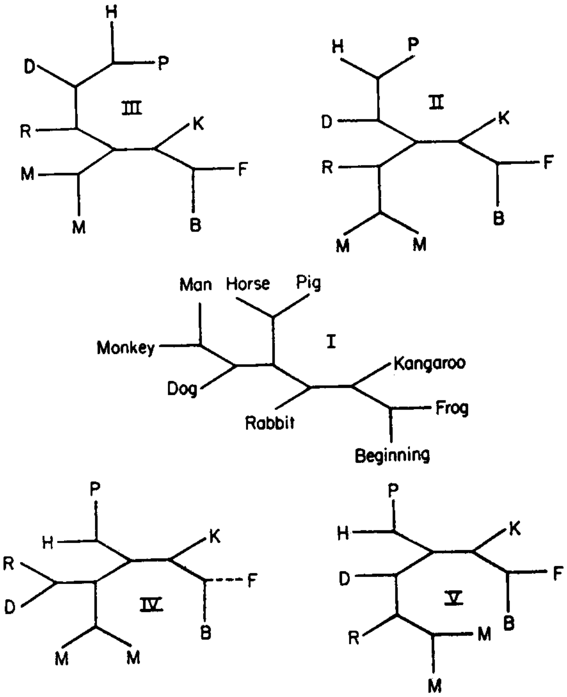

FIG. 6. Five possible vertibrate evolutionary relationship obtained from molecular sequence data. Tree I is the "best" biological tree compatible with the cytochrome C obtained by Beyer *et al.* (1974). Trees II and III were obtained from the cytochrome data by Fitch and Margoliash (1967) and Dayhoff (1972) respectively. Trees IV and V were both obtained by Dayhoff (1972) from the hemoglobin sequence data, IV for the apha chain and V for beta. See Table 1 for the nni relationships between these five trees.

TABLE 1

The values of ρ *(nni) among evolutionary trees depicted in Fig.* 6.

	I	II	III	IV	V
I	0				
II	2/M(9)	0			
III	2/M(9)	1/M(9)	0		
IV	2/M(9)	2/M(9)	2/M(9)	0	
V	2/M(9)	1/M(9)	2/M(9)	1/M(9)	0

The values of ρ *(nni) among the hemopoiesis theories depicted in Fig.* 7.

Unitarian	Dualistic	Trialistic
0		
4/M(8)	0	
5/M(8)	2/M(8)	0
U	D	T

A second example, is the comparison of the traditional theories of hemopoiesis given in Fig. 7. It is of some curiosity that the unlabeled tree structures are identical, while the differences between labeled trees are quite large. The large distance between the polyphyletic and monophyletic theories should suggest an investigation into the possibility of clear experimentally distinguishable differences.

The proposed nearest neighbor interchange metric has two important properties for biological problems. The first of these is that it yields a metric

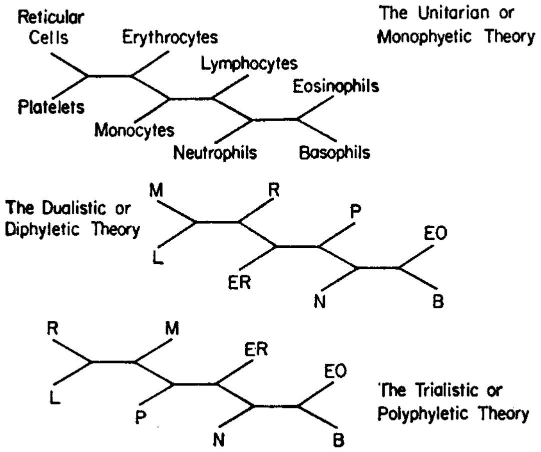

FIG. 7. The dendrographic representation of the three traditional theories of Hemopoiesis, blood cell consanguinity. (Greep and Weiss, 1973; Arey, 1968).

space and gives the distance between two dendrograms. The second is that what is being measured is the minimum number of differences in decision ordering between two tree structures. Its major disadvantage at this time is that there is no efficient algorithm known to compute the distance. Hopefully the closest partitions metric can be shown to be identical to the nearest neighbor interchange metric.

The authors wish to thank Halldor Egilsson for his help in constructing the embryonic and hemopoiesis dendrograms from the rather descriptive literature.

REFERENCES

AREY, L. B. (1968). *Human Histology*. Philadelphia: W. B. Saunders.

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

DAYHOFF, M. O. (1972). *Atlas of Protein Sequence and Structure*. Vol. 5 N.B.R. U.S.

DOBSON, A. J. (1975). *Combinatorial Mathematics III. Lecture Notes in Mathematics* Vol. **452**, 95. Berlin, Heidelberg, New York, Springer-Verlag.

FARRIS, J. S. (1973). *Syst. Zool.* **22**, 50.

FITCH, W. M., & MARGOLIASH, E. (1967). *Science* **155**, 279.

GREEP, R. O., & WEISS, L. (1973). *Histology*. New York: McGraw-Hill.

HARARY, F. (1969). *Graph Theory*. Reading, MA-Menlo Park, CA-London: Addison Wesley Publishing Company.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. New York, London, Sidney, Toronto: John Wiley & Sons.

JARDINE, N. & SIBSON, R. (1971). *Mathematical Taxonomy*. New York, London, Sidney, Toronto: John Wiley & Sons.

JOHNSON, S. C. (1967). *Psychometrika* **32**, 241.

KNUTH, D. E. (1969). *The Art of Computer Programming, Volume 1/Fundamental Algorithms*. Reading MA-Menlo Park, CA-London: Addison Wesley Publishing Company.

KORFHAGE, R. R. (1974). *Discrete Computational Structures*. New York, London: Academic Press.

MOORE, G. W., GOODMAN, M., & BARNABAS, J. (1973). *J. theor. Biol.* **38**, 423.

PHIPPS, J. B. (1971). *Syst. Zool.* **20**, 306.

PRATHER, R. E. (1976). *Discrete Mathematical Structures for Computer Science*. Boston: Houghton Mifflin.

REVERBERI, G. (1971). *Experimental Embryology of Marine and Fresh-Water Invertebrates*. Amsterdam: North-Holland Publishing Company.

SNEATH, P. H. A., & SOKAL, R. R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman & Co.

WATERMAN, M. S., SMITH, T. F., SINGH, M., & BEYER, W. A. (1977). *J. theor. Biol.* **63**, 199.

WILLIAMS, W. T. & CLIFFORD, H. T. (1971). *Taxon.* **20**, 519.

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset