Journal of
Molecular Evolution
© Springer-Verlag 1981

A Non-Sequential Method for Constructing Trees and Hierarchical Classifications

Walter M. Fitch

Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin 53706, USA

Summary. A procedure is presented that forms an unrooted tree-like structure from a matrix of pairwise differences. The tree is not formed a portion at a time, as methods now in use generally do, but is formed en toto without intervening estimates of branch lengths. The method is based on a relaxed additivity (four-point metric) constraint. From the tree, a classification may be formed.

Key words: Classifications – Trees – Additivity – Matrix methods – Neighborliness

Introduction

It is frequently the task of investigators to represent a collection of items or categories in a simple but structured fashion that shows how those items may be related to each other. The items or categories may be the languages of linguistics, psychological traits, economic factors, medical diseases, anthropological races, ecological niches, or botanical species, among others. Such structures are commonly adduced starting from a matrix that contains a measure of the relationship among the items. The degree of relationship may, broadly, be either one of similarity or dissimilarity since one is transformable into the other. There are many ways of producing such a structure (Sneath and Sokal 1973, Hartigan 1975). They are all sequential procedures in the sense that the structure is built in a piece-meal fashion with, 1, the various items being added one at a time to the growing structure in agglomerative procedures or, 2, the amorphous whole being successively subdivided by some divisive procedure, both on the strength of one or more criteria that varies from procedure to procedure. I shall detail a procedure that is unique in that it produces a structure in whole, rather than sequentially.

The structure that results from the different procedures may take many forms, but I shall restrict myself to a tree structure in which the items to be classified are at the tips of branching structures such as those shown in Fig. 1. As a biologist interested in evolution, my favorite trees are phylogenetic trees and genealogies. Such trees are normally rooted, that is, some point on the tree is marked to show the historical origin or common ancestor of the items at the tree tips. The procedure I shall present does not require that such a root exist and provides no information on its location if it does exist. The method simply provides a tip-labelled, unrooted, bifurcating branching structure of relationship, i.e., a labelled topology. Placement of a root, if one is desired, is independent of the creation of the branching structure and is left to the investigator. Branch lengths may be assigned to the segments of the tree to depict the distances between the objects under study but that is also a separate task not discussed herein and readers are free to use their own favorite method for that purpose as well. A classification scheme inheres in the tip-labelled topology if some rooting point, arbitrarily or otherwise, is selcted. That transformation of a tree to a classification is also left to the reader.

While the method is general in that it can be used on data of diverse origin, this procedure was initially motivated by a general desire to understand evolutionary realtionships. It may therefore be particularly suited for data that resulted from a branching or Markov process. Moreover, to aid in understanding the method, I shall make its exemplification concrete by using specific biological terminology and examples. In particular, I shall be interested in the relationships among a set of animals, the taxa (categories) that are to be placed at the tips of the unrooted topology. The measure of

0022–2844/81/0018/0030/$01.60

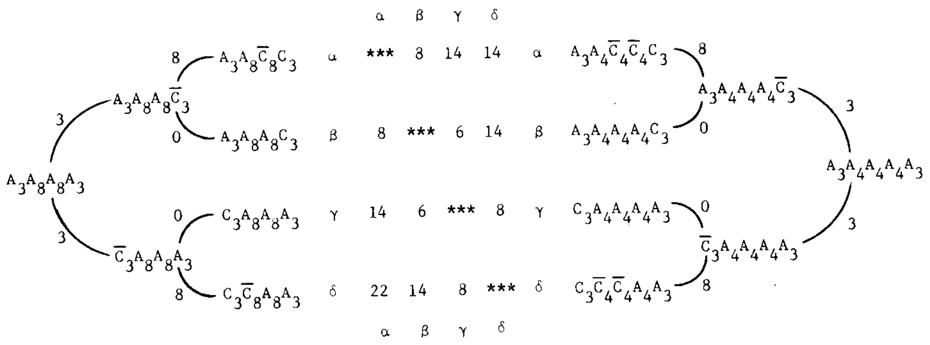

α β γ δ
8 $A_3A_8\bar{C}_8C_3$ α *** 8 14 14 α $A_3A_4\bar{C}_4\bar{C}_4C_3$ 8
3 $A_3A_8A_8\bar{C}_3$ $A_3A_4A_4A_4\bar{C}_3$ 3
0 $A_3A_8A_8C_3$ β 8 *** 6 14 β $A_3A_4A_4A_4C_3$ 0
$A_3A_8A_8A_3$ $A_3A_4A_4A_4A_3$
0 $C_3A_8A_8A_3$ γ 14 6 *** 8 γ $C_3A_4A_4A_4A_3$ 0
3 $\bar{C}_3A_8A_8A_3$ $\bar{C}_3A_4A_4A_4A_3$ 3
8 $C_3\bar{C}_8A_8A_3$ δ 22 14 8 *** δ $C_3\bar{C}_4\bar{C}_4A_4A_3$ 8
α β γ δ

Fig. 1. To the left and right of the asterisks are shown two hypothetical evolutionary trees leading to four sequences shown at the branch tips. The number and branch location of the changes (A → C) are the same in both trees but their location causes the number of observable differences between the α and δ sequences to differ so that the table of differences on the left is additive while those on the right are not

relationship between two taxa will be a distance conceived as the number of differences between some genes from those two taxa, the differences arising from changes that occurred in the lineages since the taxa had a common ancestor.

Method

Example. Consider the part of Fig. ,1 to the left of the asterisks. It is intended to depict a hypothetical ancestral gene sequence of 22 A's evolving with changes to give rise to four descendent sequences (α, β, γ and δ) all of which differ, one from another and are the source of our measure of distance. The amount of change that occurred in each line of descent is shown on the tree.

The four sequences at the tips may be examined pairwise and the number of differences in the corresponding (homologous) positions counted. The result is shown in the lower left hand portion of the matrix. The problem posed by this paper is how one might recover the branching topology of the tree given only such a matrix. This may occur when the differences are detected as melting point lowerings (DNA) or immunologically (proteins).

To the right of the asterisks in Fig. 1 is shown a different example which has, nevertheless, the identical amount of change in the various branches of the tree. The matrix of pairwise differences is not the same however. In particular the amount of observed difference between a and δ, $d(a,\delta)$, equals 14 rather than 22. It is instructive to understand the basis for the difference.

The data on the left of Fig. 1 are perfect or "additive" data in that if one were to determine a phylectic or path distance, p(i,j), for any pair of taxa i and j, by summing the values along the branches of the tree required to connect i to j, one would find that p(i,j) = d(i,j) for all i and j. If real data were to be so kind to be of this form, our task of discovering the underlying topology would be trivial because additive data conform to the additive condition (Dobson 1974).

$$d(A, B) + d(C, D) \leqslant d(A, C) + d(B, D) = d(A, D) + d(B, C) \tag{1}$$

for some labelling of the tree topology. Note that there are only three distinct ways of assigning the taxa, a, β, γ, and δ, to the four tips of the tree, remembering that reflections and rotations about the nodes where three branches meet do not alter the topological relationships of the taxa. Those three possible trees each have two pairs of tree neighbors[1] where neighbors are defined as two taxa separated on the tree but a single node and two branches. The three parts of the additive condition may be viewed as representing the sum of distances between the two neighbor pairs of the three possible ways of producing neighbors by assigning taxa to the tips of the tree. For the left-hand data of Fig. 1, the additive condition is shown as $8 + 8 < 14 + 14 = 22 + 6$. Since the two 8 values are from $d(a, \beta)$ and $d(\gamma, \delta)$, it is clear that the correct labelling of the topology requires that a and β and that γ and δ be neighbors. This is the "historical truth" we were seeking. Since this additive condition is true for additive data for every subset of four taxa regardless of the total number of taxa in the data set, the repeated application of the additive property will, with one exception, unambiously resolve the topology into a strictly bifurcating labelled tree. The one exception is when all three sums are equal. In that case, all four taxa are equally neighbors. Said differently, the branch connecting the two pairs of neighbors is of length zero.

What are the conditions required for an evolutionary process to produce additive data? Let us define each position in a sequence, such as seen in Fig. 1, as a character and the specific letter (A, C, G, T or whatever) that occurs in that position as its character state. Additive data for the most parsimoniously evolved topology result if and only if no character changes its state more than once. Unfortunately, reality perversely refuses to provide us with data from which a matrix of additive distances derives.

[1] This material was first presented at the annual meeting of the Classification Society, Boulder, Colorado, June 3, 1980. There, Douglas Carroll, to whom I am indebted, called my attention to the paper by Sattath and Tversky (1977), which also examined four objects at a time within these same "additive" terms and it is from their paper that the term neighbor derives

Let us now examine the non-additive data on the righthand side of Fig. 1. The causes of the non-additivity are the parallel changes of the central four characters from state A to state C in the descent to α and to δ. The result is that eight of the actual changes between α and δ produce no apparent sequence differences, the four central characters having identical derived character states in both taxa. One can also readily illustrate non-additivity using changes of character states that reverse the original change (say C to A) or which make a change to a new character state (say A to G or C to G).

If we now examine the resultant matrix of differences for the sums of the three possible pairs of neighbor distances we discover that $8 + 8 < 6 + 14 < 14 + 14$. This proves the data are non-additive since the two right-hand sums are not equal. However, we can see that the "historical truth" is still identifiable in that the left-hand sum still identifies α and β as neighbors as well as γ and δ. We therefore pose the relaxed additivity condition.

$$d(A, B) + d(C, D) \leqslant d(A, C) + d(B, D) \leqslant d(A, D) + d(B, C). \qquad (2)$$

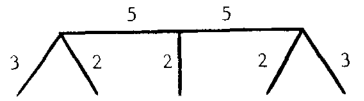
We then define, for any four taxa and their six associated pairwise distances, that those two pairs of taxa that must be equivalenced to A, B, and C, D in order to satisfy the relaxed additivity condition, have one unit of neighborliness each. We will use this unit subsequently to quantify degrees of neighborliness when more taxa will require more complicated tree structures.

Note that in order to misidentify the true neighbors, there must be more than twice as many parallel (or back) changes in two of the branches leading to two non-neighbor taxa as there are differences between the nodal sequences of the true neighbors because the two right-hand sums of equation 1 differ from the left-hand sum by exactly twice the length of the interior interval. In Fig. 1 there are six differences between the nodal sequences so there would have to have been more than 12 parallel character state changes in the descent to α and δ to get $d(\alpha, \delta) + d(\beta, \gamma) < d(\alpha, \beta) + d(\gamma, \delta)$ and hence misidentify the true neighbors. If parallel and reverse changes of character state were forbidden and only changes to new character states allowed, there would have to be more than 24 such changes before misidentification of neighbors occurred. Moreover, there would have to be even more than 12 parallel changes if there are additional parallel changes involving the neighbor pairs of taxa.

Modification of the Method. The conditions required for misidentifying neighbors seem sufficiently extreme that neighborliness might well be a powerful means of deciding tree structure. The conditions become less extreme, of course, as the number of differences between the interior nodal sequences become less and the number of character state changes in the four branches connecting the interior nodal sequences to the tips become greater, but the overall power of this approach would seem to be a great improvement over current methods.

To encompass problems of the type created by multiple character state changes in a single character, we modify the neigborliness unit to allow for the case where two specific pairs of taxa can't be assigned to the A, B and C, D of the relaxed additivity constraint because the data give the relation $d(A, B) + d(C, D) = d(A, C) + d(B, D) < d(A, D) + d(B, C)$. In such a case the two units of neigborliness must be distributed equally among (A, B), (C, D), (A, C) and (B, D), or one-half each. If the data conform to the relaxed additivity constraint in the form $d(A, B) + d(C, D) = d(A, C) + d(B, D) = d(A, D) + d(B, C)$, then all six pairs share equally one-sixth each of the two units of neighborliness.

Using All Taxa; Ideal Cases. We now approach the problem of treating more than four taxa in the same data set. Consider Fig. 2, with its five taxa, and the tree showing their relationship. There are five possible ways of selecting a subset of four taxa. For each subset of four, consider the tree pruned down to contain no taxa except the selected four and no branches except those required to connect them and determine which pairs are neighbors. We shall find that α, β are neighbors all three times that they are both present in the subset of four (the same is true for δ, ϵ), α, γ (and β, γ) are neighbors only that one time when δ and ϵ are the other members of the four-set while γ, δ (and γ, ϵ) are neighbors only that one time when α and β are the



	α	β	γ	δ	ε	
α		5	10	15	16	α
β	5		9	14	15	β
γ	10	4		9	10	γ
δ	15	14	4		5	δ
ε	16	15	10	5		ε
	α	β	γ	δ	ε	

Fig. 2. A hypothetical reconstruction of changes along the branches of a tree that gives phyletic distances (upper right of matrix) that are never less than the original distances (lower left of matrix). The length of the tree is minimal and its topology matches the neighborliness values (shown in the upper right half of Table 1) obtainable from the original distances

Table 1. In the upper half of the matrix are the neighborliness values for the interior distances in the lower half of Fig. 2

	α	β	γ	δ	ϵ
α	**	3	1	0	0
β	0	**	1	0	0
γ	5	5	**	1	1
δ	10	10	5	**	3
ϵ	10	10	5	0	**

other members of the four-set. Counting up the 10 units from those five ways would give the results shown in the upper right half of Table 1. These numbers are the ideal neighborliness values for the tree in Fig. 2, that is, the values one gets by inspection of the tree as we just found. Now examine the distance data in the lower left of the matrix and reexamine the same five four-sets of taxa, assigning neighborliness units in accord with the relaxed additivity condition (eq'n 2). Again the result is that in the upper right half of Table 1. Thus the data imply the tree structure via the relaxed additivity constraint. This is an instructive case because the distance data do not even obey the triangle inequality for the sets $[\alpha, \beta, \gamma]$, $[\beta, \gamma, \delta]$ and $[\gamma, \delta, \epsilon]$ since, for example, $d(\alpha, \beta) + d(\beta, \gamma) < d(\alpha, \gamma)$. Thus the data are not even a simple metric yet the neighborliness units lead to the intended tree whose phyletic distances are shown in the upper half of the Fig. 2 matrix, illustrating a robustness even to large perturbations of the sort that nature often introduces into real data.

In the general case for t taxa, one must examine $t!/[(t-4)!4!] = s$ sets of four taxa with 2s units of neighborliness being distributed among the various pairs of taxa. The maximum neighborliness any pair of taxa can have is $(t-2)\ (t-3)/2$ since that is the number of ways of selecting the other two taxa from the remaining $t-2$ taxa. At least two distinct pairs of taxa must have *ideal* neighborliness values that large since any tree structure must have at least two taxa (and at most t/2) separated by a single interior node. Each taxon can be paired with every other taxon in $t-1$ ways and the neighborliness values for any taxon over its $t-1$ possible neighbors is always the integral value of 4s/t.

Non-ideal Neighborliness Values. Despite the robustness of neighborliness, real data may nevertheless yield neighborliness values that do not ideally match any particular tree. An example of such data is shown in Table 2 where, in the lower left half of the matrix, are immunoglobulin distances described by Sarich (1969) for seven carnivores and a monkey. In the upper right half of the matrix are the neighborliness values for those distances. What tree then is closest to the neighborliness values?

One might determine the ideal neighborliness values for the various possible trees to see which ideal values most closely correspond to those found. One might sum the absolute differences between the observed and ideal values and conclude that tree is best for which the sum is least. The best tree by that criterion is shown in Fig. 3. The sum of the absolute differences is 28. The second best tree, which interchanges the dog and the raccoon, and thus making the dog and bear nearest neighbors, gives a sum of absolute differences of 30. That sum for the third best is 46 and involves interchanging the mink with the sea lion-seal group.

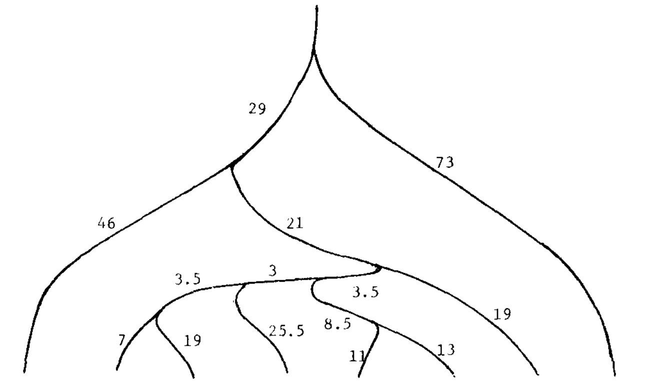
The advantage of this criterion is its simplicity, namely minimizing the sum of the errors where each two errors means that some subset of four taxa improperly identified two pairs of taxa as neighbors with respect to the putative tree. The disadvantage is that, for large numbers of taxa, it is not clear how one can be sure one has tested all the alternative tree structures that are reasonable candidates for being closest to the data. To solve this problem I now introduce a novel unbiased procedure for adjusting the observational data to make them more additive in character. This requires an examination of the length of the interior branches connecting the two pairs of neighbor taxa.

Table 2. Immunological distances[a] and neighborliness values for eight mammalian[b] taxa

	Raccoon	Bear	Dog	Seal	Sea Lion	Mink	Cat	Monkey
Raccoon		13.0	9.0	2.5	1.5	6.0	2.0	1.0
Bear	26		12.0	3.0	2.0	2.0	1.0	2.0
Dog	48	32		3.5	3.0	2.0	1.5	4.0
Seal	44	29	50		15.0	4.0	3.0	4.0
Sea Lion	44	33	48	24		6.5	3.0	4.0
Mink	42	34	51	44	38		9.5	5.0
Cat	92	84	98	89	90	86		15.0
Monkey	152	136	148	142	142	142	148	

[a] Data from Sarich (1969)

[b] Taxa are: Raccoon = *Procyon lotor*; (black) bear = *Ursus americanus*; (domestic) dog = *Canis familiaris*; (harbor) seal = *Phoca vitulina richardii*; (California) sea lion = *Zalophus californicus*; mink = *Mustela vison*; (domestic) cat = *Felis domestica*; (night) monkey = *Aotus trivergatus*. The lower left half of the matrix contains the immunological distances, the upper right half the neighborliness values for those distances



	Cat	Bear	Raccoon	Dog	Sea Lion	Seal	Mink	Monkey
Cat	***	84	96	99	90	92	86	148
Bear	84	***	26	36	33	45	36	140
Raccoon	92	26	***	48	45	47	48	152
Dog	98	32	48	***	48	50	51	155
Sea Lion	86	33	44	48	***	24	42	146
Seal	89	29	44	50	24	***	44	148
Mink	90	34	42	51	38	44	***	142
Monkey	148	136	152	148	142	142	142	***

Fig. 3. Proposed phylogeny of seven carnivores with the monkey as the out-group. Topology is that obtained from neighborliness values using the internal distances. Phyletic distances are shown in the upper right and the original immunological distances in the lower left. The topology is identical to that proposed by Farris (1972) although its length, at 282, which is truly minimal, is three units less than his tree. The scientific names of these taxa may be found in Table 2

Interior Branch Lengths. If the data were truly additive, then we could determine exactly the length (distance) associated with the interior branches. For example, the interior branch length in Fig. 1, which is 6, is necessarily one-half the difference between one of the two right-hand sums in the additivity condition (eq'n 1) and the left-hand sum, in this case (28–16)/2.

Under the relaxed additivity condition (eq'n 2) we may take as the estimate of that branch length, one-half the difference between the largest and smallest sums.[2] In the right-hand side of Fig. 1, the sums are 16, 20 and 39 so that this estimation procedure gives us the "true" value in this case, but in realistic data this cannot be guaranteed. It is true in this case because those changes producing non-additivity lie in the terminal branches (a necessary but not sufficient condition).

Now consider the problem in Fig. 2. If we wish to determine the connectiing branch length for a particular non-neighbor pair of taxa, say α/ϵ, it is sufficient to examine the data in Table 1 for the four-set $\alpha\beta\delta\epsilon$. That leads to a value of (30–10)/2 = 10, which happens to be "correct" because I chose the branch lengths to fit the data in Table 1 in a way that would minimize the length of the branches.

It needs to be recognized, however, that choosing the $\alpha\beta\delta\epsilon$ four-set was based on knowing the topology. In the realistic case, it is the topology we wish to discover. Consequently we cannot know beforehand which other two taxa to associate with α/ϵ. The practical solution is to take all possible other pairs of taxa and choose the largest of the various possible interior branch lengths. When this is done for the other two possible four-sets containing α and ϵ, namely $\alpha\beta\gamma\delta$ and $\alpha\gamma\delta\epsilon$, we obtain (25–15)/2 = 5 both times. These values prove to be the left and right halves of the overall interior distance between α and ϵ and illustrate the reason for taking the largest value observed from among all four-sets in which the two taxa are non-neighbors. These values of 5 are the largest calculated branch lengths for α/γ and γ/ϵ where α/γ and γ/ϵ are non-neighbors. *The largest calculated interior banch length will henceforth be called the interior distance* to avoid any implication whether one or more branch lengths are involved.

If the process is repeated for the taxa α and β, we discover that they are always neighbors and hence the interior distance for them proves to be zero. The same is true for δ and ϵ.

If one repeats the process for every t(t–1)/2 possible pairs of taxa, the results are as shown in the lower left half of Table 1. Note that if the 2's and 3's on the terminal branches were changed to zero, these interior distance estimates match the actual distances on the tree, but this is not always the case. Where they are unequal, the difference can result from the occurrence of multiple changes in a character or the failure of the data to meet the triangle inequality, $d(A, C) \leq d(A, B) + d(B, C)$. The triangle condition is necessarily met if the additive, four-point, condition (eq'n 1) is met.

That such pruning is useful can be seen by the carnivore immunoglobulin data of Table 2 where the neighborliness values do not correspond to any specific

[2] For Figs. 2, 3 and 4, the lengths of the branches were in fact determined using Linear Programming (hence are guaranteed to be minimal), not by using the estimating method described here

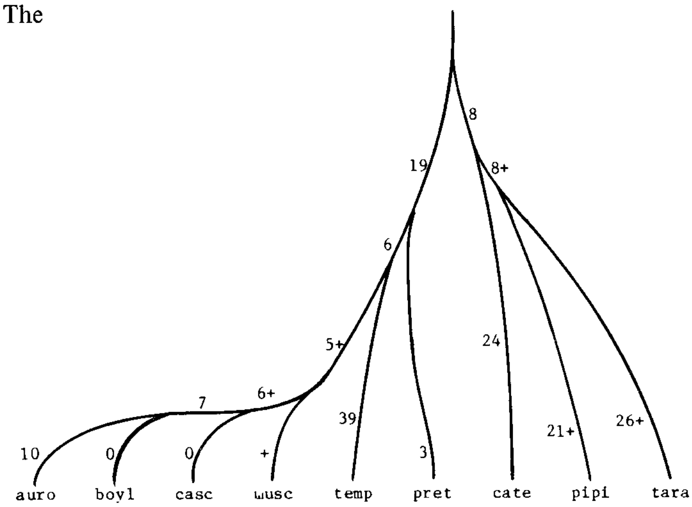
Table 3. Internal distances and neighborliness values for eight mammalian taxa

	Raccoon	Bear	Dog	Seal	Sea Lion	Mink	Cat	Monkey
Raccoon		15	10	3	3	2	1	1
Bear	4.0		10	3	3	2	1	1
Dog	5.0	5.0		4	4	3	2	2
Seal	13.5	13.5	13.5		15	4	3	3
Sea Lion	13.5	13.5	13.5	0.0		4	3	3
Mink	9.0	9.0	8.5	12.0	12.0		10	10
Cat	31.0	31.0	27.0	30.0	30.0	24.0		15
Monkey	31.0	31.0	27.0	30.0	30.0	24.0	0.0	

Derived from Table 2, which see

tree structure. If however, one estimates their interior distances as described, one gets the results shown in the lower left half of Table 3. They, in turn, yield neighborliness values shown in the upper right half of the table and these do correspond to a specific tree topology, namely the one shown in Fig. 3

Higher Order Interior Distances. The procedure for finding the interior distances from the original pairwise distances is capable of being repeated upon the interior distances themselves. This may be termed a second order interior distance. If the first interior distances do not produce neighborliness values in perfect accord with some tree, will the second order interior distances do so? The answer is that this has been observed to be true many times, and even when there was not such a perfect accord, the agreement was greater and repetition showed that the process was converging. The immunological data of Case (1978) for nine species of Rana frogs is an excellent example of degraded data since nearly 30 % (25 out of 84) of the ways of selecting three taxa give three distances that violate the triangle inequality. As might have been predicted, the first order interior distances did not produce neighborliness values in perfect accord with any tree but the second order interior distances, shown in Fig. 4, did. The resulting tree does not agree precisely with the tree of Case (1978) nor that of Farris et al. (1979) who reanalyzed these data. All three analyses give reasonably similar results. Given the uncertain quality of the data, it is probably not reasonable to try to make phylogenetic inferences from them. The neighborliness tree was, nevertheless, superior to the trees of the other authors by



	auro	boyl	casc	musc	temp	pret	cate	pipi	tara
R. aurora	***	0.0	5.7	7.2	10.7	11.2	37.5	48.5	48.5
R. boylii	10	***	5.7	7.2	10.7	11.2	37.5	48.5	48.5
R. cascadae	13	7	***	4.7	8.2	9.0	35.0	46.2	46.2
R. muscosa	12	7	7	***	5.2	7.7	31.0	42.2	42.2
R. temporaria	57	50	40	45	***	6.0	30.0	41.2	41.2
R. pretiosa	22	9	11	15	48	***	27.7	38.5	38.5
R. catesbeiana	86	65	54	48	85	54	***	11.2	11.2
R. pipiens	89	67	66	49	83	55	54	***	0.0
R. tarahumarae	97	72	79	67	107	60	59	48	***

Fig. 4. Neighborliness Tree for immunological distances from nine species of ranid frogs. The lower left half of the matrix contains the original immunological distances, the upper right half contains the second order internal distances. The distances on the branches of the tree were determined by a linear programming method (a+ is one half unit)

the independent criterion those authors used.[3]. Thus the procedure may produce a satisfactory answer even for difficult data.

Discussion

This procedure has a number of advantages. They include:

1) The use of the relaxed four-point condition minimizes the difficulty arising from the occurrence of multiple changes at a single location.
2) The procedure is not fundamentally influenced by differential rates of change in various parts of the tree (although the total number of changes may affect the probability of parallel and multiple changes at a site).
3) The method does not build the tree a bit at a time but gives a table of neighborliness values that contains all the information necessary to define the complete topolgy of a tree.
4) The same information determines the assignment of the taxa to the tips of the preceding topology.
5) The data need not even be a simple metric obeying the triangle inequality although a failure to do so may indicate that a clearly defined result may not be forthcoming (see, however, the comments below on iterative convergence).
6) The order in which the taxa and the data are presented to the algorithm are immaterial.

This procedure also has the disadvantage that, for large numbers of taxa, all possible ways of examining four of them may represent considerable computing time. It should be noticed, however, that any pair of taxa with neighborliness values of (t–2) (t–3)/2 are nearest neighbors and the iteration process may proceed on one less taxon in the manner of Sattath and Tversky (1977).

Two matters of concern need to be introduced here. The first matter of concern is the meaning and effect of the higher order interior distances. The meaning of the first order interior distance is obvious, it is the estimated distance between the two nodes closest to the two taxa for which it is the interior distance. But it is unclear what the second order interior distance is inside of. If one is only interested in getting a structure *per se*, then the meaning doesn't matter. The biologist, however, might be well advised at this stage of the development of neighborliness to submit any tree resulting from using higher orders of interior distance to some outside criterion of its satisfactoriness. I am myself just a bit suspicious that the repetitive process may be biased toward the formation of what the taxonomists call "stringy" trees. Stringy trees are those with few neighbor taxa, the stringiest having only two pairs of neighbors.

The second matter is that while iterative convergence seems always to occur, in the sense that eventually there will be no further change in the neighborliness values (or perhaps they will cycle?), it seems nearly certain that data can be manufactured (or even found in the real world) that upon convergence will produce neighborliness values that are in perfect accord with no tree. The extent of this potential problem is unknown at this point and must be explored.

Failure to obtain for every data set a perfect neighborliness pattern for some bifurcating tree should not be regarded as a defect either in the relaxed four-point condition nor in any procedure that attempts to use that constraint for tree construction. Suppose that the "true" distance on some interior branch is in fact zero. This is the case when the divergence is trichtomous rather than dichtomous. Where the distances actually correspond to this condition, the result will be to give neighborliness values that are the average of the expected neighborliness values for the three different bifurcating trees that could result if the trichtotomy were resolved into two dichotomies. Thus, what might seem to be a failure of the procedure to produce a result consistent with a specific bifurcating tree may simply indicate that the truth is a trifurcation or, more plausibly, that the data are not really adequate for the purpose of confidently deciding the issue.

Finally, there are many ways to convert a tree into a classification. One may choose a root by an arbitrary method. It could be at some point (including a node) that is deep within the tree or where, for biological data, the common ancestor of all the taxa is expected to be. The two nodes closest to this root (or three nodes if a node was chosen as the root and assuming that a terminal node was not chosen) determine the two (or three) largest divisions of the classification and each of them (node/division) gives rise in turn to two subdividions of the division. The process continues until the tips are reached and no further subdivision or levels of the classification are required. If this produces more levels of classification that one desires, one can always reduce their number (possibly with some loss of information) by any rule the user desires. For example, if node X gives rise to nodes A and Y, and Y gives rise in turn to nodes B and C, one could, ignoring Y, treat A, B and C as all members of the same immediate subdivision of X. This may be particularly appropriate when X and Y are in some sense very close to each other and don't deserve differentiation as different levels in the classification. Wiley (1979) gives an excellent discussion of other ways to attack this problem. The point is, this paper provides a novel and informative tree representation of a data

[3] The tree of Case differed from that in Fig. 4 by joining *R. cascadae* to *R. pretiosa* instead of as shown. Farris et al.'s tree joined *R. temporaria* to *R. pretiosa* and then the pair of them to *R. muscosa* before their joining to the *R. aurorae-R. boylii* group. Both Case (1978) and Farris et al. (1979) used the criterion of %SD (Fitch and Margoliash 1967) to judge their trees. The length of the neighborliness, Case and Farris et al. trees is 185, 189.5, and 196.5, respectively

matrix which can readily be transformed into a phylogenetic hypothesis by rooting it or into a classification in a manner most appropriate to the user's needs.

Acknowledgement: Thanks are due Mark Stoneking and Damon Smith for technical assistance. This work was supported by grant number DEB–7814197 from NSF.

References

Case SM (1978) Biochemical Systematics of Members of the Genus Rana Native to Western North America. Syst Zool 27:299–311

Dobson AJ (1974) Unrooted Trees for Numerical Taxonomy. J Appl Prob 11:32–42

Farris JS (1972) Estimating Phylogenetic Trees from Distance Matrices. Amer Natural 106:645–688

Farris JS, Kluge AG, Mickevich MF (1979) Paraphyly of the Rana boylii Species Group. Syst Zool 28:627–634

Fitch WM, Margoliash E (1967) The Construction of Phylogenetic Trees -- A Generally Applicable Method Utilizing Estimates of the Mutation Distance Obtained from Cytochrome *c* Sequences. Science 155:279–284

Hartigan JA (1975) Clustering Algorithms, Wiley and Sons, Ny

Sarich VM (1969) Pinniped Origins and the Rate of Evolution of Carnivore Albumins. Syst Zool 18:286–295

Sattath S, Tversky A (1977) Additive Similarity Trees. Psychometrika 42:319–345

Sneath PHA, Sokal RR (1973) Numerical Taxonomy, WH Freeman Co., San Francisco

Wiley EO (1979) An Annoted Linnaean Heirarchy with Comments on Natural Taxa and Competing Systems. Syst Zool 28:308–337

Received July 28, 1981