Classification and Related Methods of Data Analysis
H.H. Bock (Editor)
© Elsevier Science Publishers B.V. (North-Holland), 1988

FAMILIES OF TREES AND CONSENSUS

M.D.HENDY[1], M.A.STEEL[1], D.PENNY[2], and I.M.HENDERSON[2]

[1]Department of Mathematics and Statistics, [2]Department of Botany and Zoology, Massey University, Palmerston North, NEW ZEALAND

Some asymptotic properties of the symmetric difference metric for tree comparisons are described, settling two conjectures from [4]. These were that for binary trees the probability that the distance $d(T_i,T_j)$ equalled the maximum value of 2n-6 approached 1.0 as $[n]\rightarrow[n+1]$ $\forall n>3$. For non-binary trees this limit appeared to be less than 1.0. The metric is then applied to the analysis of evolutionary trees from sequences of macromolecules. A family of trees is considered as a subset of similar trees, such as may be generated from a tree by rearranging taxa around a short internal edge. Two applications are reported. In the first with minimal trees from different proteins the trees fall into different families, largely composed of trees from the same protein. In the second case with near minimal trees from the combined data set the trees still fell into more than one family though of very similar trees.

RECENT WORK ON THE SYMMETRIC DIFFERENCE METRIC

Programs that reconstruct evolutionary trees may produce more than one tree and metrics for comparing these trees are useful. The symmetric difference (or partition) metric [1,13] has been particularly useful because its distribution is known with binary trees for up to 16 labelled pendant points (see figure 1), or for up to 12 with non-binary trees [4]. The metric is easy to compute [2,10]. Two conjectures were made earlier [4] about the behaviour of this metric as n, the number of labelled points, increases to infinity. For two randomly selected binary trees T_i,T_j it appeared that the probability that the distance $d(T_i,T_j)$ equalled the maximum value (2n-6) approached 1.0 as n increased. For non-binary trees this limit appeared to be less than 1.

Our previous approach [4] required the enumeration of all classes of phylogenetic trees and as such is not suitable for studying large values of n. Recent work by one of us [15] has used a new approach to settle these conjectures, and has extended our knowledge of the

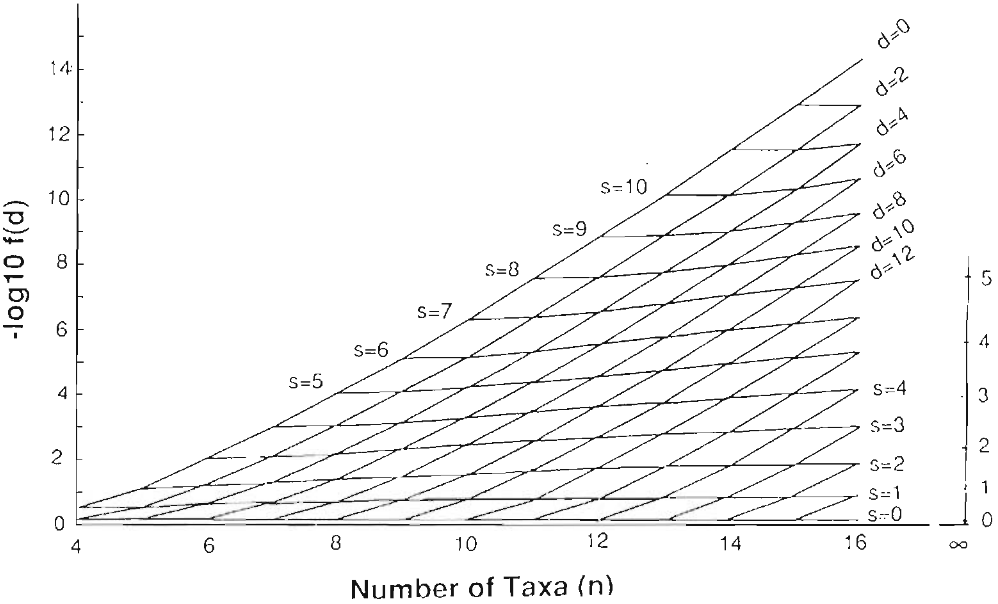

Figure 1. Properties of the Symmetric difference metric. The expected frequencies for the distance between randomly selected pairs of binary trees. For a given number of taxa n, the relative frequency f(d) is the proportion of pairs of trees (among all pairs) that are distance d apart. For d = 0,2,4,... 2n-6, -(log f(d)) is plotted. Diagonal lines connect points of the same d value and horizontal lines connect points with the same S (n-3-d/2) values. Thus, for example, the frequency f of pairs of binary trees with 11 taxa where d=6 is $10^{-4.4}$. The expected frequencies are given in table 4 of [4]. The right hand axis has asymptotic values of s from 0 to 5, calculated by the method of Steel [15].

properties of this metric.

We let BPT(n) be the set of binary trees having n labelled pendant vertices, and let $q_n{}^S$ be the proportion of all pairs $T_1,T_2 \in$ BPT(n), such that each pair has exactly S pairs of equivalent internal edges.

Alternatively, $q_n{}^S$ may be regarded as the probability that two randomly chosen binary trees T_1,T_2 (from BPT(n)) have $d(T_1,T_2) = 2(n-3-S)$.

The main result is that as $n \to \infty, q_n^S$ has a Poisson distribution with mean $\mu = 1/8$.

That is, $\lim q_n^S = e^{-1/8}/8^S S!$

In particular, $\lim q_n^0 = e^{-1/8} \approx 0.88$, settling a conjecture [4]

and $\lim q_n 1 = 1/8e^{-1/8} \approx 0.11$

Thus for large values of n, most (88%) pairs of trees have no edges in common, while 99% have at most one common edge. This makes the metric useful for hypothesis testing where trees constructed from DNA sequences may be expected to be "similar" and a metric for which most trees are 'far apart' is desirable.

Indeed the above results also show that

1) The probability that two trees in BPT(n) are distance < m apart $\to 0$ as $n \to \infty$, for any fixed m.

2) The expected distance between two binary trees tends to the maximal distance, $(2n-6)$.

It also appears that q_n^S for $S > 0$ is monotone decreasing (and it has been shown that q_n^0 is monotone increasing) so the table [4] can be extrapolated for n>16. Together with the main result above this gives a monotone convergence to the Poisson distribution.

Many of the main results above do not carry over to non-binary trees (PT(n)), suggesting that the metric works more "naturally" on binary trees than on non-binary. In particular, the expected distance between two phylogenetic trees does not tend to the maximal distance, confirming a conjecture in [4] - indeed, using a result from [3], the expected distance can be shown to converge to $1/(4\ln 2-2)-1/2 \approx 0.7943$.

Also, the probability that two randomly chosen trees $T_i, T_j \in PT(n)$ (the set of all phylogenetic trees, binary and non-binary, with n labelled pendant vertices) are a maximal distance apart tends to zero as $n \to \infty$ (unlike the case for BPT(n) where the limit was $e^{-1/8}$). Basically this is because the trees most distant from any $T \in PT(n)$ are binary trees and $|BPT(n)| / |PT(n)| \to 0$ as $n \to \infty$. Additional details are in Steel [15].

Table 1. Distances between the trees in figure 2, only the entries for the first 16 trees are shown. Edge lengths were given a value equal to their percentage of the total length of the tree. The weighted version of the symmetric difference tree comparison method was used[12].

2	13.0														
3	12.0	13.0													
4	22.0	34.0	33.0												
5	34.0	20.0	33.0	13.0											
6	35.0	35.0	21.0	13.0	14.0										
7	14.0	31.0	29.0	34.0	48.0	50.0									
8	25.0	15.0	26.0	43.0	30.0	45.0	14.0								
9	25.0	28.0	14.0	43.0	43.0	32.0	14.0	12.0							
10	15.0	28.0	27.0	23.0	36.0	37.0	29.0	39.0	39.0						
11	26.0	12.0	25.0	34.0	20.0	35.0	43.0	25.0	38.0	14.0					
12	25.0	25.0	12.0	33.0	33.0	21.0	41.0	36.0	24.0	13.0	13.0				
13	10.0	26.0	24.0	29.0	42.0	44.0	10.0	22.0	22.0	24.0	37.0	35.0			
14	22.0	12.0	23.0	39.0	26.0	41.0	24.0	9.0	21.0	35.0	21.0	32.0	13.0		
15	27.0	44.0	42.0	53.0	67.0	70.0	19.0	33.0	33.0	50.0	64.0	63.0	25.0	39.0	
16	13.0	27.0	26.0	39.0	52.0	54.0	35.0	47.0	47.0	34.0	46.0	46.0	29.0	42.0	32.0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

APPLICATIONS TO PROBLEMS WITH EVOLUTIONARY TREES

The decision on whether or not to use a consensus tree, and if so what form of consensus tree [8], requires careful thought. For example, in one study we had 31 minimal trees for 5 proteins from 11 mammals. The trees from different proteins are very similar [9]. Nevertheless, the different proteins are giving slightly different trees, probably because of parallel changes or reversions and more study is needed to understand the differences.

For a set of minimal trees derived from more than one protein the consensus may not be close to any of the minimal trees and could give a poor estimate of the 'true' phylogeny. In such cases a consensus from each protein may be more useful than a single consensus tree. Choosing just a single consensus tree may ignore information in the data.

We need to obtain more information from the trees to understand the differences between minimal trees from different proteins. If the tree representing the true phylogeny has one or more short edges (edges with few nucleotide changes) then taxa can be rearranged around these short edges at little cost (cost being an increase in length of the tree).

Such a set of trees we call a 'family', all trees in the family are only a small distance (on the tree comparison metric) from all other members. We need to be able to recognise whether a set of trees can be usefully represented as one or more families. In this context, a family is defined as all trees within a fixed distance of a tree T (based on the symmetric difference) and in this respect it is similar to a clique.

We considered four possibilities for the results of the analysis.

1) All near minimal trees, from both individual proteins and from the combined sequences, form a single family. This would happen if the sequences were all giving a consistent answer, apart from one or more short edges on the tree.

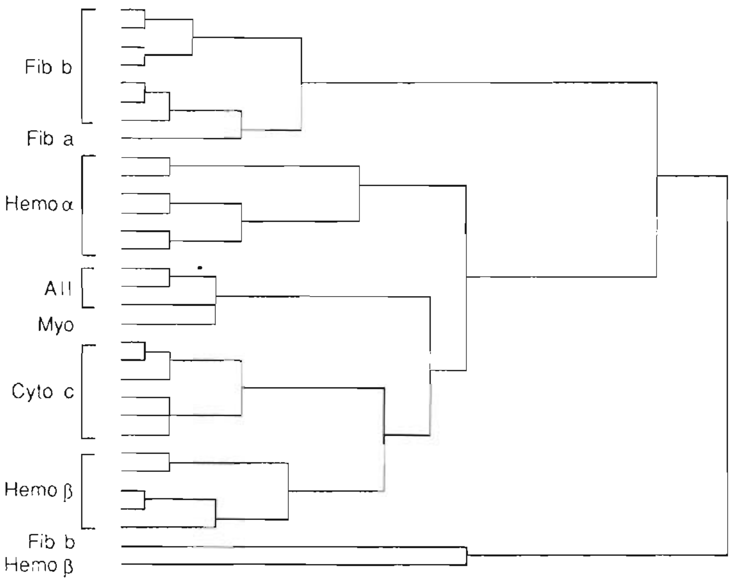

Figure 2 Cluster analysis of minimal trees. Minimal trees (and for some proteins trees one step longer than minimal) were collected for 6 proteins, and for the combined protein sequences. The edge lengths were weighted [6] and the weighted comparisons made with the symmetric difference tree comparison metric [12]. The resulting distances between trees were clustered using complete linkage clustering [14]. Trees are identified as being derived from cytochrome c (Cc), hemoglobin α (hα) and β (hβ), fibrinopetides A and B (Fa and Fb), myoblobin (myo), or from the combined sequences (all).

2) Near minimal trees for each protein form their own family, near minimal trees are more similar to each other than to near minimal trees from other proteins.

3) More than one family occurs and trees from each protein fall into the different families.

4) The concept of families of trees is not useful for analysing these results. There is no pattern in the differences between the near minimal trees.

The data is 6 protein sequences for the same 11 mammalian taxa which have been converted to 'best guess' nucleotide sequences [9,11]. From these sequences (separately and combined) the 31 minimal, or near minimal, trees have been found and are shown in [9] as trees 12-39 and in figure 2 of [11]. (In the following, 'near minimal' includes minimal trees.) We used a branch and bound method [5] to find all minimal trees (and in some cases all trees close to minimal).

RESULTS

Using the 6 sequences (separately and concatenated) the expected length of each edge was calculated [5]. Weighted comparisons [11] of the 31 trees were made and are shown in Table 1. Complete linkage cluster analysis was selected as the main clustering approach and the results shown in figure 2. This gives the interesting result that, in

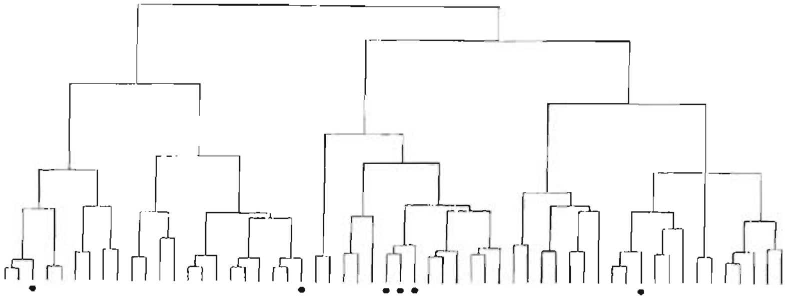

Figure 3. Cluster analysis of minimal and near minimal trees. The 56 shortest trees from the combined sequences of 7 proteins clustered using complete linkage. Those marked • are the six shortest trees (using weighted lengths [6]), they fall into four families of trees.

general, trees from each protein are clustered together. The result is consistent whether single linkage, average linkage, of Ward's clustering method are used. A difference is that one tree from the hemoglobin β is separated from other hemoglobin β trees when complete linkage is used. The other anomolous tree (a fibrinopeptide B tree) is isolated with all the clustering methods used.

Our interpretation of this result is that the trees from each protein deviate from the real tree, but that the different proteins deviate in different ways. Examples would be one protein separating rabbit and rodent whereas other sequences place them together. Another protein may lead to a different 'error' on the tree. Such a result is to be expected in an evolutionary tree with a stochastic mechanism of change. There is no evidence for the different proteins supporting several distant trees, that would be difficult to explain on an evolutionary mechanism. In this case looking for 'families' appears to have found more information about the relationships between the trees than taking a simple consensus tree.

Another example compares the near minimal (including minimal) for seven proteins (cytochrome c, hemoglobin α and β, fibrinopeptides A and B, myoglobin [11] together with α-crystallin. There are 56 minimal and near minimal trees with lengths (with linear weighting [11], from 109.47 to 112.39 (there are then no trees with lengths between this value and 113.50).

These 56 trees have been compared with weighted edge lengths and clustered by the same procedures (figure 3). All the trees are quite similar but the results have still not converged to a single family. They are consistent with 4 families of trees and each family includes one of the shortest trees (<110, marked • in figure 3). The trees most similar to the shortest trees are only slightly longer, the more distant trees are considerably longer (within the range of lengths specified. This is depicted in figure 4 where the height of the surface is the complement of weighted tree length and the horizontal axes are an ordination of the tree comparison metric using multidimensional scaling (SPSS-X). Similar trees are close together in the X-Y plane and so isolated peaks in the surface represent natural families of near minimal trees.

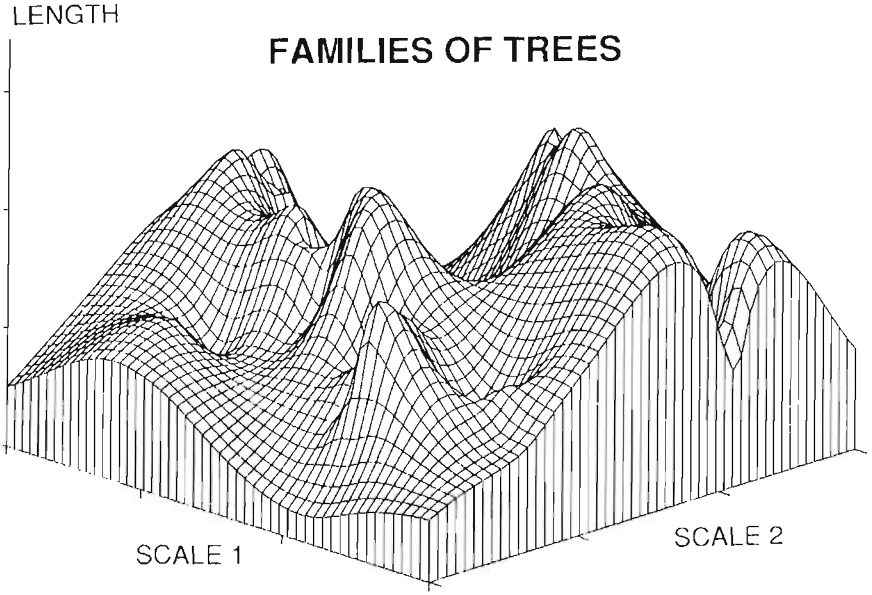

Figure 4. A graphic representation of the concept of families of trees. Peaks on the surface represent the shortest trees in the set of 56 trees from figure 3. Similarities between trees based on the symmetric difference metric are represented by an ordinate in the X-Y plane. "Scale 1" and "Scale 2" are the 2 principle axes from a multidimensional scaling ordination of the tree comparison matrix (r^2 for the two axes is 67%).

DISCUSSION

Because there is still more than one family of tree with the combined proteins, then it is premature to select just one of them. The simplest explanation is that there is insufficient sequences for the near minimal trees found by this method to have converged to a single family. In particular, it is known [10] that dog is the least stable taxon on different trees and it is desirable to add a second carnivore sequence to see whether this will reduce the differences between the trees. Other explanations are possible and should not be overlooked. For example, parallel evolution on different lines of descent should be checked for. However, just to build a single consensus tree from these results glosses over important information on the reasons for the differences.

Our main interest has been in evaluating the use of trees as predictors of new optimal tree that are formed as more data becomes available. In such cases a binary tree may be more accurate than a non-binary tree. However, in a study on the classification of a group it is probably better to take a consensus tree that is non-binary. This would give more stability to the classification, though may not give quite as good predictions about the optimal tree when more data is available. Thus the choice on the form of consensus tree to be used will depend on the purpose of the study being undertaken, in this case, the biological problem to be solved.

Keywords: Consensus trees, evolutionary trees, families of trees, protein sequences, symmetric difference, tree comparison metrics.

REFERENCES

[1] Bourque,M., Arbres de Steiner et reseaux dont varie l'emplagement de certains sommets, Ph.D. thesis, Universite de Montreal, Quebec, Canada. (1978)

[2] Day,W.H.E., Optimal algorithms for comparing trees with labelled leaves. J.Classification. 2 (1985) 7-28.

[3] Foulds,L.R. and Robinson,R.W., Enumeration of phylogenetic trees without points of degree two. Ars Comb. 17A (1984) 169-183.

[4] Hendy,M.D., Little,C.H.C. and Penny,D., Comparing trees with pendant vertices labelled. S.I.A.M. J.Appl.Math. 44 (1984) 1054-1065.

[5] Hendy,M.D. and Penny,D., Branch and bound algorithms to determine minimal evolutionary trees. Math.Biosc. 59 (1982) 277-290.

[6] Hendy,M.D., and Penny.D., Edge lengths of trees from sequence data. Math.Biosc. 83 (1987) 157-165.

[7] Kruskal,J.B. and Wish.M., Multidimensional scaling. Sage Publications, Beverley Hills. 1978

[8] Margush,T. and McMorris.F.R., Consensus n-trees. Bull.Math.Biol. 43 (1981) 239-244.

[9] Penny,D., Foulds,L.R. and Hendy,M.D., Testing the theory of evolution by comparing phylogenetic trees constructed from 5 different protein sequences. Nature 297 (1982) 197-200.

[10] Penny,D. and Hendy,M.D., The use of tree comparison metrics. Syst. Zool. 34 (1985) 75-82.

[11] Penny,D., and Hendy,M.D., Estimating the reliability of evolutionary trees. Molec.Biol.Evol. 3 (1986) 403-417.

[12] Robinson.D.F., and Foulds,L.R., Comparison on weighted labelled trees. Pages 119-126 in Lecture Notes in Mathematics. Vol 748 (1979) Springer-Verlag, Berlin.

[13] Robinson.D.F., and Foulds,L.R. Comparison of phylogenetic trees. Math. Biosc. 53 (1981) 131-147.

[14] Sneath,P.H.A. and Sokal,R.R., Numerical Taxonomy. Freeman, San Fransisco. (1973)

[15] Steel,M.A. Distribution of the symmetric difference metric on phylogenetic trees. (1987) (manuscript)