Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences

David Penny*, L. R. Foulds† & M. D. Hendy‡

* Department of Botany and Zoology, and ‡ Department of Mathematics and Statistics, Massey University, Palmerston North, New Zealand
† Operations Research, University of Canterbury, Christchurch, New Zealand

The theory of evolution predicts that similar phylogenetic trees should be obtained from different sets of character data. We have tested this prediction using sequence data for 5 proteins from 11 species. Our results are consistent with the theory of evolution.

THE theory of evolution continues to be a focus for nearly all biological thought. Nevertheless, there have been doubts about the philosophical status of the theory, particularly on the extent to which it can be tested or falsified. The best known of these doubts have been expressed by the leading philosopher on scientific method, Karl Popper[1], who concluded that "darwinism is not a testable scientific theory, but a 'metaphysical research program'—a possible framework for testable scientific theories". Popper did not in any way reject evolution. He pointed out (ref. 1, p. 169) that "no serious competitor has come forward" and commented on "the strange similarity between my theory of the growth of knowledge and darwinism". In particular, he suggests that only one prediction is possible from darwinism: "Gradualism is thus, from a logical point of view, the central prediction of the theory. (It seems to me that it is the only prediction.)" (ref. 1, p. 172).

Popper has recently modified these criticisms[2], pointing out that criticisms by some authors were inconsistent, such as that natural selection is a tautology, and that it explains too much. A tautology explains nothing, so cannot simultaneously explain "too much". In addition, the suggestion was made that the existence of an evolutionary tree was falsifiable, but no reasoning was given for this new opinion.

These criticisms have aroused considerable interest[3-5]. Most of the discussion has, however, been qualitative, so it would be useful to be able to make quantitative tests. This is the purpose of the present article.

Popper makes the important distinction between the existence of an evolutionary tree (an evolutionary history) and the mechanism of evolution put forward to explain the processes that produced that history. Here we present a programme, applying graph theory[6,7], by which it is theoretically possible to refute the existence of an evolutionary tree.

Another prediction from evolution

It has been long been considered that protein sequence data contain evolutionary information[8]. In particular, the tree of minimal length (minimum number of mutations, maximum parsimony) makes no assumptions about the mechanism of evolution, and has been widely used as a model for evolutionary relationships[9-11]. Given any comparative data, irrespective of origin, one can construct trees of an evolutionary form, and hence a minimal tree must exist. So in general, finding a minimal tree for protein sequence data is not in itself independent evidence for the existence of an evolutionary tree.

However, another prediction can be made if there has been an evolutionary tree and if the maximum parsimony model is a good predictor of that tree. The prediction is that minimal trees with the same taxa should be similar, or 'congruent'[12], when constructed from different protein sequences. We need a measure of tree similarity having biological significance so that the values can be compared with the hypothesis that the trees are randomly selected. Such a measure is described below.

Our strategy is to take different protein sequences for a common set of taxa, find all the minimal (and near minimal) evolutionary trees and then compare them. Should the probability be high that these trees are unrelated, this would indicate that the protein sequences do not contain similar evolutionary information, and hence would contradict the existence of an evolutionary tree for those taxa.

We conclude that the existence of an evolutionary tree for these taxa is a falsifiable hypothesis, and therefore meets Popper's criteria for scientific theories. In addition, our methods allow us to identify a consensus tree which incorporates the most common features of all the near minimal trees.

Finding minimal evolutionary trees

It is easily shown that any minimal evolutionary tree can be represented by a binary tree. Using the double factorial notation (!!) there are $(2n-5)!! = 1 \times 3 \times 5 \ldots \times (2n-5)$ unrooted binary trees spanning n sequences[13,14]. To determine the trees of maximum parsimony (minimal length), one must potentially consider a vast number of different tree topologies. But for all but a small number of sequences ($n \leq 8$), present day computers cannot consider them all[10].

Table 1 The frequency $p(m, n)$ of occurrence of $d(Ti, Tj) = m$ for binary trees spanning n taxa $4 \leq n \leq 11$

	$m = 0$	2	4	6	8	10	12	14	16	$E(n)$
$n = 4$	3.3×10^{-1}	6.7×10^{-1}								1.34
5	6.7×10^{-2}	2.7×10^{-1}	6.7×10^{-1}							3.20
6	9.5×10^{-3}	5.7×10^{-2}	2.4×10^{-1}	7.0×10^{-1}						5.24
7	1.1×10^{-3}	8.5×10^{-3}	4.6×10^{-2}	2.2×10^{-1}	7.3×10^{-1}					7.32
8	9.6×10^{-5}	9.6×10^{-4}	6.5×10^{-3}	3.8×10^{-2}	2.0×10^{-1}	7.5×10^{-1}				9.40
9	7.4×10^{-6}	8.9×10^{-5}	7.0×10^{-4}	4.8×10^{-3}	3.1×10^{-2}	1.9×10^{-1}	7.7×10^{-1}			11.46
10	4.9×10^{-7}	6.9×10^{-6}	6.2×10^{-5}	4.8×10^{-4}	3.6×10^{-3}	2.7×10^{-2}	1.8×10^{-1}	7.9×10^{-1}		13.52
11	2.9×10^{-8}	4.6×10^{-7}	4.7×10^{-6}	4.0×10^{-5}	3.3×10^{-4}	2.8×10^{-3}	2.3×10^{-2}	1.7×10^{-1}	8.1×10^{-1}	15.55

$E(n) = \Sigma\, mp(m, n)$ is the weighted mean or 'expected value' of $d(Ti, Tj)$ for randomly selected binary trees Ti, Tj spanning n taxa.

0028–0836/82/200197—06$01.00

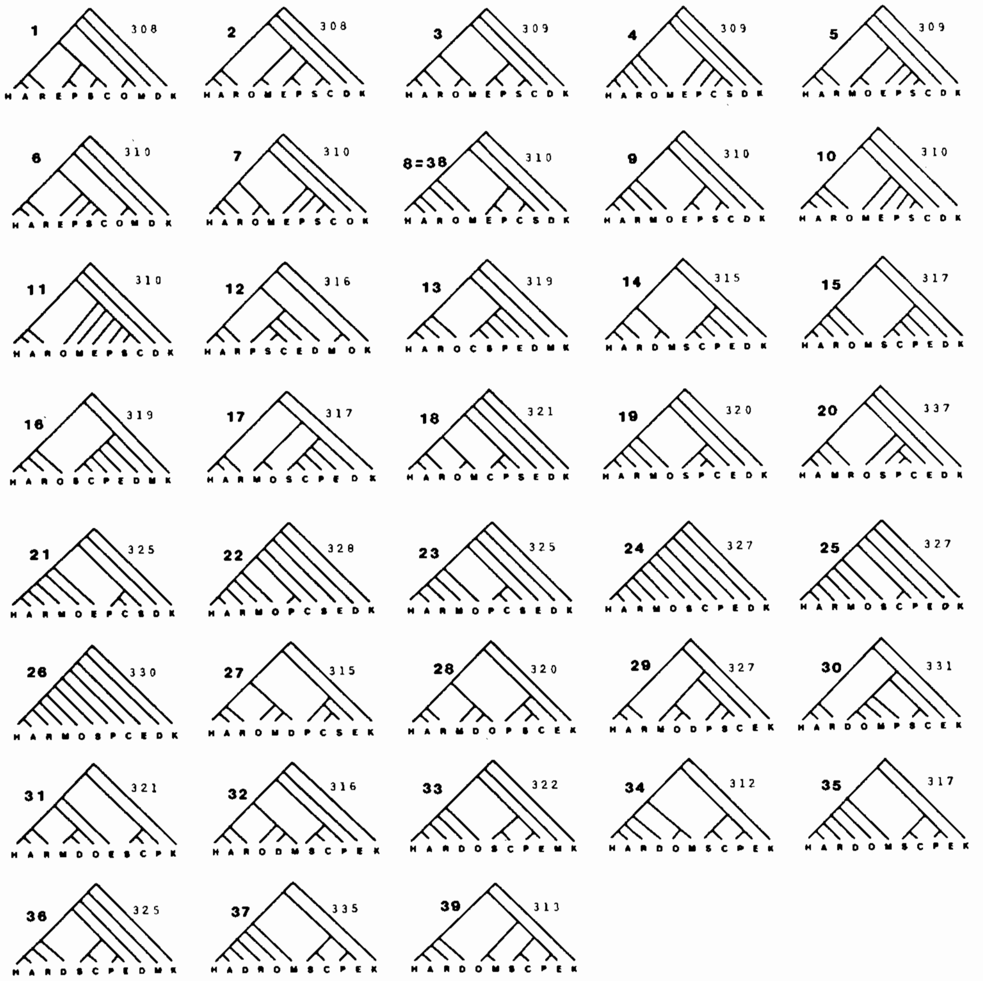

Fig. 1 The 39 minimal and near minimal trees spanning the 11 taxa given in Table 2. $T1$–$T11$ are generated from the complete sequences, $T12$–$T17$ from cytochrome *c*, $T18$ from fibrinopeptide A, $T19$–$T26$ from fibrinopeptide B, $T27$–$T32$ from haemoglobin *A*, and $T33$–$T39$ from haemoglobin *B*. The trees are arranged with their sequences in increasing lengths (see Table 3). The total number of mutations of each tree for the combined sequences is shown. The lengths of individual links are not shown, only the branching sequence is indicated. See Table 2 legend for definition of letter abbreviations.

We have shown that the problem of finding minimal length trees is an example of the Steiner problem in graphs[15,16] which makes no assumptions about the existence of an evolutionary history. Two recent techniques developed by the authors (refs 15–17 and D.P., in preparation) have enabled us to solve this for some larger sets of sequences (in one case a tree with 25 species has been proved minimal). These techniques rely on the inherent structural content of the data and are found not to work well on random data. This is an indication that there may be considerable tree-like structure within the sequence data. However, these observations are not easily quantifiable and the inability to use such methods on some data sets of this size ($8<n<25$) could not be taken as evidence for or against the existence of an evolutionary tree.

Here we have used a recently developed (ref. 17 and D.P., in preparation) 'branch and bound' method[7] which, for the data in Table 2, has found all minimal and near minimal trees.

Comparison of trees

Several methods have been proposed[18] that measure some features of the difference between a pair of trees, but it is important that any method chosen has biological relevance. Apart from satisfying a historical curiosity, a major application of an evolutionary tree is to assist in the classification of the taxa into groups at different levels. The set of taxa that are descendants of a given common ancestor will form a group of related taxa. In a strictly bifurcating (binary) tree spanning n taxa, there will be $n-2$ such non-trivial classes of the taxa. As in Robinson and Foulds[19], we measure the difference $\mathrm{d}(T1, T2)$ of two trees $T1$ and $T2$, as the number of classes which are derived from $T1$ or $T2$, but not both. It is easily shown that this measure forms a metric on the space of all trees spanning these n taxa.

Given a non-directed tree T (one in which the root has not been specified), the removal of any internal link will partition the taxa into two subsets, one of which (depending on orientation) will be a group of related taxa. Waterman and Smith[20] have shown that the set of all such partitions uniquely defines T. Thus, $\mathrm{d}(T1, T2)$ also represents the number of partitions of the taxa formed by deleting internal links, which differ in the two trees. This analysis is independent of whether or not the tree is a rooted tree.

Probabilities and tree comparisons

We can, for any specific tree T, determine the value of $\mathrm{d}(T, T')$ for each of the $(2n-5)!!$ trees T'. Then the proportion of trees with $\mathrm{d}(T, T')=m$ will represent the probability that a randomly selected tree T' is distance m from T.

For binary trees the number of internal links is $n-3$, so each tree defines $n-3$ partitions. If Ti, Tj have m partitions in common, $\mathrm{d}(Ti, Tj)=2(n-3-m)$ which is even, and can range from 0 to $2n-6$. The value $\mathrm{d}(Ti, Tj)=0$ can occur only for $i=j$, so the frequency of $\mathrm{d}(Ti, Tj)=0$ is $1/N$, where $N=(2n-5)!!$. If $\mathrm{d}(Ti, Tj)=2$ then there is only one link at which the two trees disagree. If we delete one internal link of Ti, there are two alternative ways of rejoining it to give values $\mathrm{d}(Ti, Tj)=2$. This could occur at any of the $n-3$ internal links, so we find $\mathrm{d}(Ti, Tj)=2$ with frequency $(2n-6)/N$.

The values of $\mathrm{d}(Ti, Tj)$ over all pairs of binary trees spanning n species for $4 \leqslant n \leqslant 11$ have been determined using recursive generating functions (D.P. and M.D.H., in preparation). These results are summarized in Table 1.

We can, for a given value of m, use these frequencies to estimate the probability of randomly selecting two trees Ti, Tj with $\mathrm{d}(Ti, Tj)=m$. Referring to the case $n=11$, we would, for example, expect $\mathrm{d}(Ti, Tj)=4$ to occur with probability

Table 2 Nucleotide sequences derived from the 5 polypeptides for 11 taxa

Taxon	Haemoglobin A	Haemoglobin B
R	CUCGGGGGGAUUACGGAAACUAGGAAAGGCAGCUGCAGAG	GCCACAAACACCCCCAUCUCCCGGCGGGACCAACCACAAGCUGACUAACAGCA
S	GAGGGGGGCAUUACGGGCAAAAGCCGGAGACUGUAACAGA	AACAGGCACGCUCCGAUCACGCGACGAUACAAAACGGCAGCGUGAUCAGUCAU
E	GACGGGGGGAUUGCGCAAACUAGCCGGGGACGUUAACGAG	GCGUGGCUUGCCCAGCUCAACAGGCCAUGAGCACCAGCGGCGUGAUACGACCA
K	GGGGAGGGGACAACACAAGCAAAACGGAAAAGCUGAAGGC	GCCAGAAACAGCAAACCAAAGAGCCGUGGCAAAAAAAAAAAAUAAUAAGAACG
M	GUCAGGAGGAUCACGGAAGAACCCCGGGGCAUCAGCCGAG	GCAUGGCUCGGCCCGAUCUCGCGGGACGACCAACCAUCGAAAUACCACGCGCU
O	CACAGAAGGAUUGAACAAAAAAGCCGGGACAAAUUCAGAA	GCGAAUCACGCCCAGAUCUCGCAACGCGAACAGACAAAGGCAUACUAACAGCA
D	CACAAAAGGCCAACGCAAAACACCCGGGGCAUGAACAGGC	GCCAGUCUCGGCCCACUCACCCAGGAAUACCAAAAAAAAGCUGACUAACAGCA
P	GGCGGGGCGAUUGAGCGAAAAAGCAGGGGCAUCUAACAAA	GCCUGGCUUGGCCCGAUCAAGCGGCGAUAACAAACAAAGGAGUGAUCCGUCCA
H	CACGGGGCGAUUACGGAAAAACCGAAAGGCAGCUGCAGAG	GCCACUCACGCCCCGAUCACCCGGCGGGACCGCCCAACGGCUGACUAACCGCA
C	GGGGGGGGCAUUACGGGCAAAAACCGGGGAAUCUAGCGAA	AACAGGCACGGUACGACAAAGCAACGAUACACAACGCAGGCGUGACCAGUCAU
A	CACGGGGCGCUUACGGAAAAACCGAAAGGCAGCUGCAGAG	GCCACUCACGCCCCGAUCACCCGGCGGGACCGCCCAACAGCUGACUAACCGCA

Taxon	Fibrinopeptides A and B		Cytochrome *c*
R	GAGGGAGGACCCG	AGAGAGAAGCCUAG	AAUUCAUCCCUAC
S	GCGUGGGGAACCG	AGGGCAACUCCUGA	GCAGCUUACACGA
E	CAGAGAGGAACAG	AGGACAAGUACUGA	GCAGCUAACACCA
K	CAAAGAGACAACG	AGGGUAAAGAGGGA	GCAGCUAAUCCGC
M	GAGGACAGAAAAC	AGCAAGUGAAUUAG	GCAGGUUACCCGC
O	GCGGGAAACAAAC	AGCGAGAGUUUCGA	GCAGGUUACCCAC
D	AUAAGAGGAAACG	AGGAUGUAGACGGA	GCAGCUUACACGC
P	GGCAAAGGAACCG	AGGCCAAGUCAGGA	GCAGCUUACACGA
H	GUGGGAGGACCCG	UAAGAGAGGUUUAG	AAUUCAUCUCUAC
C	GCCCAGGGACCCG	AGGCCAAGUGGUGG	GCAGCUUACACGA
A	GUGGGAGGACCCG	UAAGAGAGGUUUAG	AAUUCAUCUCUAC

R = Rhesus monkey (*Macaca mulatta*); S = sheep (*Ovis ammon*); E = horse (*Equus caballus*); K = kangaroo (*Macropus conguru* or *Macropus giganteus*); M = mouse (*Mus musculus* or *Rattus rattus*); O = rabbit (*Oryctolagus cuniculus*); D = dog (*Canis familiaris*); P = pig (*Sus scrufa*); H = human (*Homo sapiens*); C = cow (*Bos primogenius*); A = ape (*Pan troglodytes* or *Gorilla gorilla*).

4.7×10^{-6}. The weighted values, $E(n)$, given in Table 1 are the estimate of the expected value $(2n-6)$. Small values of $d(Ti, Tj)$ are very rare.

We have at this point established: (1) a method that can guarantee to find all minimal and near minimal trees, (2) a quantitative method of comparing trees and (3) a method of associating frequencies with the comparison of trees. The next stage is to apply these methods to comparative biological data.

Application to sequence data

When this study began there were polypeptide sequences available that were common to 11 taxa. The proteins are cytochrome *c*, haemoglobin A, haemoglobin B, fibrinopeptide A and the last 13 amino acids of fibrinopeptide B. The original data are from the *Atlas of Protein Sequence and Structure*[21], together with its supplements. There are cases, such as kangaroo, where two species of the same genus have been used (*Macropus conguru* and *Macropus giganteus*), but this does not affect the result. Table 2 lists the species. Our methods will work with more species and/or more sequences (for example, myoglobin, α-crystallin A, RNase), but we frequently find just one sequence is not available. There is an urgent need for more coordination in selecting sequences for analysis.

The protein sequences were converted to nucleotide sequences[22] (with haemoglobin B, some nucleotide sequences were available[23]). The sequence data were edited to remove invariant sites and other sites with no comparative information, as these have no effect on the structure of the phylogenetic trees (refs 10, 16 and D.P., in preparation). This left only a small number of sites, particularly for cytochrome *c*, as it shows little variation among these taxa. The final data are listed in Table 2, with 40 sites from haemoglobin A, 43 from haemoglobin B, 13 from fibrinopeptide A, 14 from fibrinopeptide B and 13 sites from cytochrome *c*.

Evolutionary tree construction

For 11 taxa there are 17!! = 34459425 unrooted binary trees to be considered. The branch and bound algorithm was applied to each of the five protein sequences individually, as well as to the combined sequences. Table 3 lists the numbers of trees close to minimal length for each of the five data sets and for the combined data. The 39 trees whose lengths are within 1.25% of the minimal lengths were selected for detailed comparisons and are illustrated in Fig. 1. In order that they should be presented as rooted evolutionary trees, the marsupial (kangaroo) was selected as determining the root or ancestor of the tree.

Tree comparisons

It can be seen from Fig. 1 that there are only two identical trees among these 39, $T8 = T38$. Using the comparison algorithm outlined above, we obtain the ${}_{39}C_2 = 741$ values of $d(Ti, Tj)$ ranging from 0 (1 value) to 14 (8 values) out of the maximum of 16, with a mean value of 7.57. Interpolating from the frequencies in Table 1, we would expect such a value to occur between a pair of random trees with probability 1.9×10^{-4}. There are, of course, 741 comparisons but they cannot be considered to be independent values. The average similarity for comparisons of trees from the same sequences is 4.1 and for comparisons of trees from different sequences is 8.33. The expected and observed values of *m* are as follows:

m =	0	2	4	6	8	10	12	14	16
Exp.	0	0	0	0	0	2	17	125	597
Obs.	1	53	87	163	200	145	84	8	0

Table 3 The minimal length and number of trees close to minimal length for each of the five sequences individually and combined

Sequence	Minimal length	min	(min + 1)	(min + 2)	(min + 3)	(min + 4)
Cytochrome *c*	17	6*	55	195	321	368
Fibrinopeptide A	29	1*	37	403	2,724	12,449
Fibrinopeptide B	36	8*	126	475	1,313	4,660
Haemoglobin A	89	1*	5*	26	143	400
Haemoglobin B	124	3*	4*	25	41	105
Combined sequences	308	2*	3*	6*	0*	8

* Those 39 trees with length $\leqslant$1.25% of the minimal. These 39 trees were used in the comparative analysis and are shown in Fig. 1.

Table 4 The average distance between trees derived from two sequences

	CS	Cc	FA	FB	HA	HB
CS	2.0/3.9	7.0	8.0	9.8	8.0	5.3
Cc	5.8	3.1	8.3	10.2	6.0	5.7
FA	6.9	8.3	—	4.8	10.0	11.3
FB	8.2	10.2	4.8	4.5	8.8	11.8
HA	8.4	7.6	10.3	11.2	—/4.7	6.7
HB	6.5	7.6	11.1	11.5	9.0	2.7/4.4

The values in the upper triangle are from comparisons among the 21 minimal trees, the values in the lower triangle are from comparisons among the 39 trees studied. CS = combined sequences, Cc = cytochrome *c*; FA, FB = fibrinopeptides A and B; HA, HB = haemoglobins A and B.

There is thus a strong divergence away from random towards the trees being very similar. Note that it is logically possible for the trees to have been more dissimilar than expected.

Table 4 gives the mean values of d(Ti, Tj) between trees of different sequences, the upper values being obtained only from the minimal trees, and the lower values obtained from all 39 trees. The largest mean that occurs is 11.75 between the minimal length trees of fibrinopeptide B and haemoglobin B. This value would occur between a pair of random trees with probability 1.8×10^{-2}. All other values have probabilities less than this, ranging down to 1.1×10^{-5} for fibrinopeptides A and B.

Clearly, we can reject any idea that the trees from the different sequences are independent. The different protein sequences give trees that are markedly similar, showing a relationship between them that is consistent with the theory of evolution. This supports the theory but of course does not prove it; scientific theories are falsifiable but not provable[1,24].

Consensus tree

Of the 34459425 unrooted binary trees for 11 taxa, we have selected the 39 near minimal trees for further study. The minimal trees with the combined sequences are one estimate for the most likely tree from these data, but it is also possible to derive a 'consensus tree' which incorporates the most common features of the 39 trees.

We can, by using the partitions, find a particular tree T such that the sum $\mathrm{d}(T, T1)+\mathrm{d}(T, T2)+\cdots+\mathrm{d}(T, T39)$ is minimal. In calculating the tree distances, we computed the number of times each particular partition occurred in a given tree. The nine most frequent partitions and the number of trees in which they occur are {H, A} (39); {H, A, R} (37); {S, C} (30); {S, C, P, E} (28); {S, C, P} (22); {K, D} (21); {H, A, R, O, M} (16); {E, P} (12); {O, M} (12) (see Table 2 legend for definition of abbreviations). For example, the first partition is {H, A} which is human and ape, and this occurs in all 39 trees. Each partition is defined by listing the smaller of the two subsets, and so in this case ({H, A}) the remaining nine taxa are in the other subset. The tenth most frequent partition occurs in only seven trees.

The partition {E, P} with frequency 12 is inconsistent with partition {S, C, P} with frequency 22. If we delete {E, P}, the remaining 8 partitions uniquely define the 8 links of a binary tree T spanning the 11 taxa. This tree is, in fact, the tree $T7$ in Fig. 1 of length 310 on the complete data and with an average distance of 5.5 to the other trees. We call this tree the consensus tree of the set of trees and it could be regarded as being representative of the set. It is the tree which will give the minimal sum of m values when comparisons are made with all 39 trees. It is a markedly better tree on this test than any other. Other consensus tree methods are available[25] but were less appropriate for this application.

The consensus tree does have the advantage over the two combined sequence minimal trees ($T1$, $T2$) in that in the ungulates, the horse (perissodactyl) is separated from the three artiodactyls (cow, sheep and pig). References to results of mammalian phylogeny from palaeontological evidence can be found elsewhere[26,27]. In McKenna's published scheme[26], lagomorphs (rabbit) would be the first branch after the marsupials. This does not occur in any of our 39 trees.

Conclusion

The general conclusions from the present work are that (1) it is possible to make falsifiable predictions from the hypothesis that species have been linked in the past by an evolutionary tree and (2) there is strong support from these five sequences for the theory of evolution. There may be exceptions where different sequences will lead to different trees as, for example, in the serial symbiosis theory[28]. Also, in pre-cellular evolution a network with circuits may be a better model than a tree[29]. Note also that this work has so far been confined to the question of the existence of an evolutionary tree and has not discussed the mechanism of evolution. The work can be extended by using criteria of optimality that assume particular mechanisms of evolution.

An interesting philosophical question would arise if the results of this work had falsified the prediction that the trees would be similar. Would this disprove the theory of evolution, or could it just mean that the sequences had changed so rapidly that they had lost all information about their early history, thus contradicting the hypothesis of Zuckerkandl and Pauling[8]? It could be argued that because proteins from different species can be aligned so readily, this in itself is independent evidence that the proteins retain evolutionary information. However, it is probably true that specific predictions from hypotheses, rather than the hypotheses themselves, are falsifiable. This idea is inherent in Popper's writing, but is more clearly expressed by Lakatos[30]. To this extent, we suggest that Popper's criticisms of evolutionary theory have shown incompleteness in the application of evolutionary theory, but at the same time evolutionary theory has helped clarify some inadequacies in Popper's model of the growth of knowledge.

Received 12 October 1981; accepted 16 March 1982.

1. Popper, K. *Unended Quest: An Intellectual Autobiography* (Fontana, London, 1976).
2. Popper, K. *Dialectica* **32,** 339–355 (1978).
3. Halstead, B. *New Scientist* **87,** 215–217 (1980).
4. Ruse, M. *New Scientist* **89,** 828–830 (1981).
5. Editorial *Nature* **290,** 75–76 (1981).
6. Harary, F. *Graph Theory* (Addison-Wesley, Reading, Massachusetts, 1969).
7. Carre, B. *Graphs and Networks* (Clarendon, Oxford, 1979).
8. Zuckerkandl, E. & Pauling, L. *J. theor. Biol.* **8,** 357–366 (1965).
9. Dayhoff, M. O. & Eck, R. V. *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Springs, Maryland, 1966).
10. Fitch, W. M. *Am. Nat.* **111,** 223–257 (1977).
11. Goodman, M., Czelusniak, J., Moore, G. W. & Romero-Herrara, A. E. *Syst. Zool.* **28,** 132–163 (1979).
12. Mickevich, M. F. *Syst. Zool.* **27,** 143–158 (1978).
13. Cavalli-Sforza, L. L. & Edwards, A. W. F. *Evolution* **21,** 550–570 (1967).
14. Felsenstein, J. *Syst. Zool.* **27,** 27–33 (1978).
15. Hendy, M. D., Foulds, L. R. & Penny, D. *Math. Biosci.* **51,** 71–89 (1980).
16. Foulds, L. R. & Hendy, M. D. *J. molec. Evol.* **13,** 127–150 (1978).
17. Hendy, M. D. & Penny, D. *Math. Biosci.* **59** (in the press).
18. Smith, T. F. & Waterman, M. S. *Am. Math. Mon.* **87,** 552–553 (1980).
19. Robinson, D. F. & Foulds, L. R. *Springer Lect. Notes Math.* **748,** 119–126 (1979).
20. Waterman, M. S. & Smith, T. F. *J. theor. Biol.* **73,** 789–800 (1978).
21. Dayhoff, M. O. *Atlas of Protein Sequence and Structure 1972* (National Biomedical Research Foundation, Silver Springs, Maryland 1972).
22. Penny, D., Hendy, M. D. & Foulds, L. R. *Biochem. J.* **187,** 65–74 (1980).
23. van Ooyen, A. *et al. Science* **206,** 337–344 (1979).
24. Popper, K. R. *Objective Knowledge* (Oxford University Press, 1972).
25. Margush, T. & McMorris, F. R. *Bull. Math. Biol.* **43,** 239–244 (1981).
26. McKenna, M. C. in *Phylogeny of Primates* (eds Luckett, W. P. & Szalay, F. S.) 21–46 (Plenum, New York, 1975).
27. Szalay, F. S. in *Major Patterns in Vertebrate Evolution* (eds Hecht, M. K., Goody, P. C. & Hecht, B. M.) 315–374 (Plenum, New York, 1976).
28. Schwartz, R. M. & Dayhoff, M. O. *Science* **199,** 395–403 (1978).
29. Eigen, M & Winkler-Oswatitsch, R. *Naturwissenschaften* **68,** 217–228 (1981).
30. Lakatos, I. in *Method and Appraisal in Physical Science* (ed. Howsen, C.) 1–40 (Cambridge University Press, 1976).