9

Testing the Theory of Descent

DAVID PENNY, MICHAEL D. HENDY, AND MICHAEL A. STEEL

> If it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down (Darwin, 1859: 189).

The comment of Popper (1976:168) that "Darwinism is not a testable scientific theory, but a *metaphysical research program*—a possible framework for testable scientific theories" is sometimes used to question the scientific status of evolutionary theory. The original comment was not in any way "antievolution," and indeed, Popper noted "the strange similarity between my theory of the growth of knowledge and Darwinism" (Popper, 1976:169). Later, Popper (1978, 1984) limited these criticisms, but this has not always satisfied critics. One of our interests has been to determine, without ambiguity, if evolutionary theory could meet Popper's criteria for the demarcation of science.

The introductory quote given above is one of many examples of how Charles Darwin considered his theory of evolution to be falsifiable. Although Darwin is remembered today for the general aspects of theory of evolution, in his day-to-day research, he made many predictions from his theory—then sought (or made observations) to test the predictions. Some of the best known involve his experiments on plants, including work on orchid flowers (both in structure and function), pin and thrum flowers of primrose (and other dimorphic and trimorphic forms of flowers), and on the power of movement of plants (Ghiselin, 1969; Allan, 1977; Penny, 1985). This last example of the power of movement in plants is particularly interesting in that he had left a record of his reasoning. The first extract is from his autobiography.

> For in accordance with the principles of evolution, it was impossible to account for climbing plants having been developed in so many widely different groups, unless all kinds of plants possess some slight power of movement of an analogous kind.

And again, the delightful comment in a notebook from 1839,

> Is there any very sleepy mimosa, nearly allied to the Sensitive Plant?

(Details of sources are in Penny, 1985). The prediction that all plants should have some "slight power of movement" led to two books on his research, *The Movements and Habits of Climbing Plants* and *The Power of Movement in Plants*. Many other examples of how Darwin used predictions to direct research could be given, including features of human evolution.

It was not just a coincidence that Charles Darwin was looking for predictions from theories. During the early stages of the development of his theory (after the voyage of the Beagle), Darwin read widely in many areas of science. This was partly in order to understand what was required of a good scientific theory. His reading included works on the philosophy of science by Herschel, Whewell and Comte (Ruse, 1975; Schweber, 1977). Each of these authors emphasized the importance of prediction in science and, judging from letters to Charles Lyell (Schweber, 1977), Darwin was well aware of the importance of a good scientific theory leading to predictions.

In this century, Karl Popper (1963, 1972) has developed this theme further and emphasized the potential falsifiability of *scientific* theories. A theory is more useful (and therefore better) if it prohibits (excludes) a larger proportion of possible observations. The approach claims to be both descriptive and prescriptive. It is descriptive in that it claims to describe how the most effective scientists have worked, and it is prescriptive in that it advocates how scientists should aim to work. Hypotheses (or conjectures) are considered tools that are judged on their effectiveness in helping scientists devise new and more powerful tests whether the tests be observational, experimental, or analytical. None of this denies that sociological and cultural factors play a role in science—as long as such factors are considered descriptive of scientific procedures and not prescriptive of how scientists should select theories. During the past two decades there has been a strong anti-intellectual movement suggesting that hypotheses are largely determined by cultural factors, and consequently, to this extent are arbitrary. Such a movement has failed to develop falsifiable predictions and is not scientific by Popper's criterion for the delimitation of science.

In this chapter we review our approach to the study of evolutionary trees. This has been developed within a strong Popperian framework (Riddiford and Penny, 1984) of aiming to develop falsifiable hypotheses. After discussing some of the general issues involved, we then discuss the question of how good methods are for inferring trees, particularly from molecular data.

IS EVOLUTION A SCIENTIFIC THEORY?

Similar Trees From Different Sequences

The simple prediction from the "theory of descent" is that, because the sequences share the same tree pattern of ancestry, the optimal trees from

different sets of data should be similar. (From the proposed stochastic nature of the mechanism of mutation and selection it would be surprising if the trees were identical. Indeed, it would be more devastating to Darwinism if different sets of short sequences always gave identical trees). Comparing results from different data sets had been used previously (e.g., Mickevich, 1978) and our interest was in getting quantitative results.

Our first major project was to compare trees derived for the same 11 mammalian taxa but from different sets of sequence data. We saw three requirements for being able to test the prediction of similar trees from different sequences. These were the ability to:

1. find the optimal tree(s) for 10 or more taxa;
2. find a tree comparison measure to compare trees objectively; and
3. derive the distribution of this tree comparison measure.

Finding the Optimal Tree

Sufficient taxa were required so that it would be most unlikely to get the same tree by chance. Fortunately, sequence data for 11 taxa were available for five proteins or peptides.

For 11 taxa there are 34,459,425 (17!!) binary trees. The number of trees increases exponentially with the number of taxa; therefore, any method which considers all trees cannot be efficient. It was shown quite early in the project that searching trees for the optimal tree(s) is an example of a set of problems known to be NP-complete (Graham and Foulds, 1982). This result implies that it is most unlikely that an efficient method will ever be found for a complete search on a large number of taxa. William Day has extended this analysis to other related problems with trees (see Day et al., 1986).

Finding optimal trees requires a program which is unbiased toward any subclass of trees. For example, it could happen that a program tended to group together adjacent taxa in the data matrix. An error of this nature could lead to the trees from different sequences being more similar then expected, not through common descent, but through program limitations. A quite different problem is that an optimality criterion, for example parsimony, could tend to favor some trees by bringing together isolated taxa. This will be discussed later.

In 1980, a search through all trees for 11 taxa was estimated to require 55 days with the computers we had available. The solution to this problem was the development of a branch and bound algorithm (Hendy and Penny, 1982), which is now widely used. It reduced the computing time for these data to just under five minutes, while still guaranteeing to have found all optimal trees. This met our first objective of finding optimal trees for the five sets of data.

Finding a Tree Comparison Measure

When the parismony branch and bound program was applied to each of the five different proteins, the minimal trees were not identical, but did

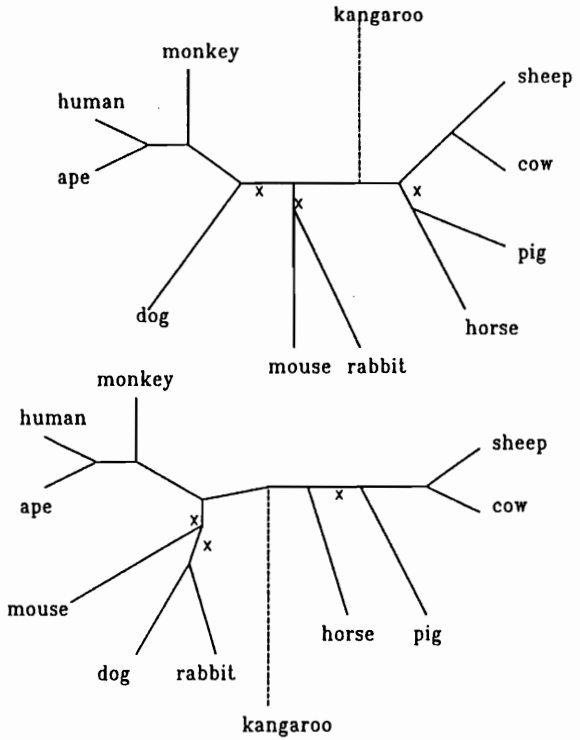


Figure 9-1 Application of symmetric tree metric. Two trees selected from the minimal trees from β-hemoglobin (top) and α-hemoglobin (bottom) sequences. On the symmetric difference metric there are six differences. These are indicated by a small cross on the three nonequivalent edges of each tree.

indeed look "similar" (Penny et al., 1982). The second objective was to find a useful tree comparison metric that would allow an objective comparison of these trees. The symmetric difference metric had been developed (Robinson and Foulds, 1981) from a parallel interest in trees and had an efficient method for its calculation (Penny and Hendy, 1985b). A more efficient method was developed by Day (1985).

The symmetric difference metric on two trees counts the number of edges that occur in one, but not both, trees. Edges are equivalent if they partition the taxa into the same two subsets (see Edge Bipartitions). This is repeated for each edge of the tree in turn. Figure 9-1 shows two minimal trees from different sequences which are distance six differences apart. The differences between the two trees are indicated by a small cross on edges which have no equivalent edge on the other tree.

Deriving the Distribution of the Tree Comparison Measure

At this point, it is unclear whether finding six differences is significant. What is the probability that two randomly selected trees would have six differences? In order to answer that question, the problem of deriving the distribution of the tree comparison metric must be solved. The expected distribution has been calculated for up to 16 taxa (Hendy et al., 1984) and is shown graphically in Penny and Hendy (1986) and Hendy et al., (1988). The distribution is highly asymmetric, which is useful for many biological applications because it is particularly sensitive for closely-related trees. We find the probability of randomly selecting trees on 11 taxa with six or fewer differences (as in Fig. 9-1) is 4×10^{-5}.

Similar results were found from comparisons of minimal trees from other sequences (Penny et al., 1982), where each pair of minimal trees from different sets of sequences was more similar than expected by chance. Thus, it is fair to claim that the original prediction that minimal-length trees from different data sets would be similar, is supported. The method of analysis allowed the possibility for the theory of descent to fail. We have more confidence in the theory if it passes quantitative tests.

Our conclusion is that the theory of descent can meet the same quantitative standards as expected in other areas of science. There is no need for "special pleading" that evolution is hard to quantify. The project led to improved techniques in several areas for studying trees. The problem of finding optimal trees was shown to be NP-complete (Graham and Foulds, 1982). Branch and bound methods were developed (Hendy and Penny, 1982) so that optimal trees could be found in reasonable time for up to at least 16 taxa. A tree comparison metric (Robinson and Foulds, 1981) was implemented by showing it could be calculated efficiently (Penny and Hendy, 1985b). The expected distribution of this metric was derived (Hendy et al., 1984) so that quantitative tests can be made. The analysis of a complex scientific problem in order to get falsifiable predictions can be a productive approach to science.

Additional Predictions

Are Shorter Trees Better?

The parsimony criterion for an optimal tree assumes that shorter trees (those requiring fewer changes) are better estimates of evolution than longer trees. As a corollary of this we would expect that trees shorter for one data set should also be shorter on other data sets. This is different from the previous section which compared minimal trees from independent data sets.

A test of this prediction is shown in Figure 9-2 where trees requiring 124 to 133 changes on the β-hemoglobin data were selected. The lengths of these trees were then determined on the combined sequences from cytochrome *c*, fibrinopeptides A and B, and α-hemoglobin.

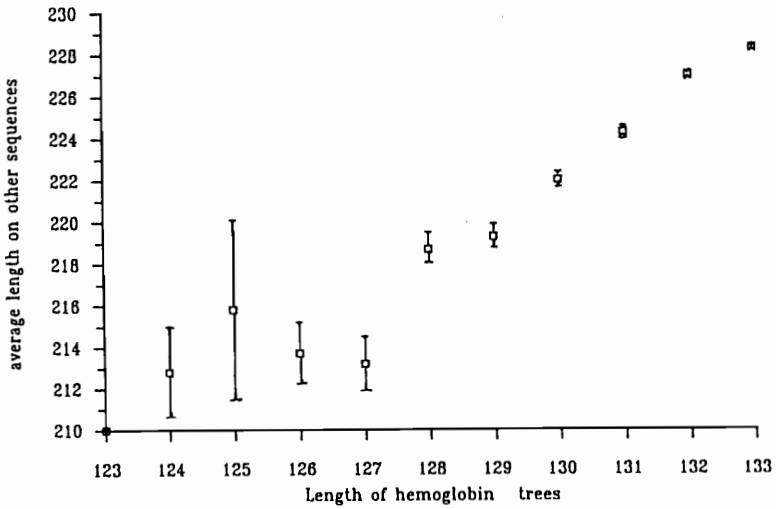


Figure 9-2 Lengths of trees with β-hemoglobin and other sequences. The shortest trees for β-hemoglobin sequences were selected. The lengths of these trees were then recalculated on the combined sequences from cytochrome c, fibrinopeptides A and B, and α-hemoglobin. *On average*, trees that require fewer mutations with β-hemoglobin, require fewer mutations with the new sequences. The bars are twice the standard error of the mean. (Adapted from Penny and Hendy, 1985a.)

The results in Figure 9-2 show that, *on average*, trees requiring fewer mutations with β-hemoglobin require fewer mutations with other sequence data. The test was repeated for each of the other four sequences (Penny and Hendy, 1985a), giving even better results. Evolution is a stochastic process, and the shortest tree on an individual sequence cannot be guaranteed correct.

Sampling Error and Convergence

It is generally recognized that the sequences used in a particular study may be too short to give an accurate prediction. If the only problem is that the sequences are too short, then it is expected that the optimal tree should become a better estimate as the sequences become longer. We cannot, of course, measure this difference directly. However, what can be measured is whether optimal trees from different data sets become more similar as sequences become longer.

Perhaps the best method for selecting subsets of data is by the random resampling of columns. This can be done by either bootstrapping or jackknifing. In bootstrapping (Felsenstein, 1985; Penny and Hendy, 1985a), subsets of columns from the data matrix are randomly selected *with* replacement. These subsets can have the same number of columns as the original data. A column may be omitted from a particular subset, or selected more than once. With jackknifing (Penny and Hendy, 1985a, 1986), the subsets are randomly selected *without* replacement. Consequently, the subsets must be shorter than the original sequence. Jackknifing methods can also be divided into those where subsets may overlap and those with disjoint subsets (where no column occurs in both subsets). In this latter group, which we call hobbits or halflings, each subset contains no more than half the columns.

Trees can be formed from these subsets by standard methods and the results analyzed using a tree comparison metric. There have been two ways of comparing results. Felsenstein (1985) determined the internal edges that occur in at least 95% of the trees. Penny and Hendy (1985a, 1986) studied the rate of convergence as longer sequences (subsets) were used.

With either form of jackknifing we have measured the average distance between optimal trees (using the symmetric difference tree comparison metric) from each subset of columns. What we would expect is that trees from different subsets would become more similar as the subsets contained more columns. This indeed is the case as is shown in Figure 9-3. Instead of showing the value from the symmetric difference metric directly, we have converted it, using the calculated distribution of the metric (Hendy et al., 1984, 1988), to the equivalent number of trees. These results show that even the five sequences for each mammal is insufficient to allow

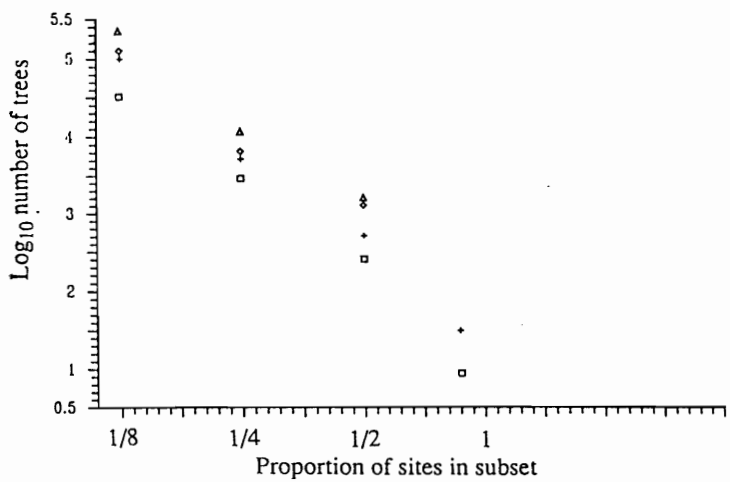


Figure 9-3 Convergence with larger subsets of columns. Minimal length trees were found for jackknife samples from the combined sequences for six proteins. The samples contained 12.5%, 25%, 50% or 92% of the columns in the data matrix. The symmetric difference metric was used to find the average distance between optimal trees from each group of subsets. Optimal trees became more similar as the samples became longer. The results allow four different methods to be compared (see Penny and Hendy, 1986 for an explanation of the symbols and further discussion). (Adapted from Penny and Hendy, 1986.)

convergence to a single tree. In the cases we have studied (Penny and Hendy, 1985a, 1986), the trees, as expected, become more similar as the subsets become larger. This is independent evidence for evolutionary information in the sequences.

These resampling approaches allow tentative answers to several interesting questions. Is it likely that a different optimal tree will be found if more information is gathered for each taxon? If so, which trees? How large a subset of trees is necessary to be confident of including the correct tree? Do some methods for inferring trees converge faster than others as longer sequences are used?

Is bootstrapping better than jackknifing with larger-and-larger samples? The answer may depend on the application. In a taxonomic study, an author may not wish to propose a new taxonomic category unless confident that future work will support it. This is using stability of a single edge of the tree as a criterion. In such a case, bootstrapping may be the preferred test since it does not accept an edge if there is reasonable doubt. If, however, the intent is to find the best estimate of the phylogeny (the tree which gives the best prediction as more data become available), then convergence may be suitable. Under these circumstances the aim may be to find how large a subset of trees is required to be confident that the subset includes the correct tree. It must be noted that although the test allows a decision as to whether convergence has occurred, it is still possible for methods to converge to an incorrect tree. This is discussed later (see Hadamard Transformations).

Testing Other Models

From a Popperian viewpoint, an important feature of the scientific approach is that scientists should be able to give rational explanations of why they use a particular theory. There is an asymmetry here in that it is not claimed that arriving at the original theory was necessarily rational—the creative process is much more complex than that. The aim is for decisions between competing ideas to be rational, and that we should be able to give conditions under which we would reject a currently favored hypothesis.

In the first section we discussed testing the prediction that trees from different sequences led to similar trees. Could we test nonevolutionary models? One well-known astronomer, Sir Fred Hoyle, had ideas of the earth being bombarded with influenza viruses from passing comets (Hoyle, 1984). We called this the unhealthy falling object (UFO) model. From the UFO theory it is not expected that viral sequences would arrive in a particular order consistent with an evolutionary tree. The details of the theories examined are described in Henderson et al. (1989).

One step of the analysis required the comparison of how well a column of data fitted a binary tree, compared to the null model of the star (or big bang) tree (Fig. 9-4). On this null method we could imagine a single ancestor from which all existing species had been independently derived. That is, no pair of species is more closely related than any other pair. Even with

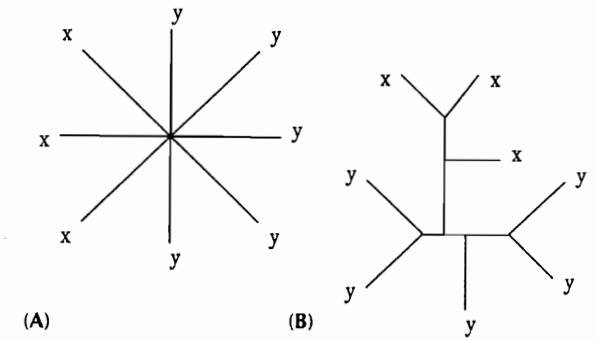


Figure 9-4 (A) The star (or big bang) tree which is a common null hypothesis (Thompson, 1975). (B) is used to illustrate that even if sequences were generated by a star tree process, it is still possible to fit the data with fewer changes to a binary tree.

data generated independently from a common origin (the star tree model) it would still fit more closely to a binary tree. For example, Figure 9-4 has a 3,5 character [three of one color (code or character state), five of the second] and this would require at least three changes on a star tree (Fig. 9-4A). But even if the star tree was correct, many binary trees could be drawn that required only one change on the binary tree. The number can be calculated from Figure 9-4B by forming a tree with eight taxa from the two subtrees with a new edge. It thus became important to know the expected number of changes required to fit a column of data to a random binary tree. If a binary tree model is a good representation of the data, then the number of changes required to best fit the data to a tree should be significantly less than the number required for data generated from the star tree model.

For 2-state colors (codes or character states) with frequencies a and b, on a single column of n taxa ($n = a + b$), we found $f_m(a,b)$ binary trees whose minimal coloring required m changes (Carter et al., 1990).

$$f_m(a,b) = (m-1)!(2n-3m)(2n-5)!!N(a,m)\mathrm{N}(b,m)/(2n-2m\text{-}1)!! \quad (1)$$

where, a is the number of pendant vertices with the first color (y in Fig. 9-4),
b is the number of pendant vertices with the second color (x in Fig. 9-4),
$a \geq b$, $a + b = n$, the number of taxa,
m (≥ 1) is the minimal number of changes on the tree,
!! is the double factorial [$(2n-5)!! = 1 \times 3 \times 5 \ldots \times (2n-5)$; $0!! = -1!! = 1$], and

$$N(n,m) = \begin{cases} (2n-m-1)!/(n-m)!(m\text{-}1)!2^{n-m} & \text{if } n \geq m, \text{ and} \\ 0 & \text{if } n < m. \end{cases}$$

Table 9-1 Values of $F_m(a,b)$ for Four to Eight Taxa*

Number of Taxa	a	b	Single Column Distribution: # Trees for m = 1,2,3,4				Weighted Average Distribution: t	#Trees for m = 1,2,3,4			
			1	2	3	4		1	2	3	4
4	2	2	1	2			z	1	2		
5	3	2	3	12			z	3	12		
6	4	2	15	90			z	12	78		
							y	3	12		
	3	3	9	54	42		z	9	45	36	
							y	0	9	6	
7	5	2	105	840			z	60	570		
							y	45	270		
	4	3	45	360	540		z	36	234	360	
							y	9	126	180	
8	6	2	945	9,450			z	360	4,680		
							y	270	2,250		
							x	270	2,250		
							w	45	270		
	5	3	315	3,150	6,930		z	180	1,440	3,420	
							y	45	855	1,620	
							x	90	720	1,710	
							w	0	135	180	
	4	4	225	2,250	5,544	2,376	z	144	1,152	2,592	1,152
							y	72	432	1,440	576
							x	0	648	1,296	576
							w	9	18	216	72

*$F_m(a,b)$ is the number of trees with m changes with a and b occurrences of two character-states. The values for the single column distribution are calculated from equation (1). For the weighted average distribution, t indicates the topology (unrooted binary tree) using the convention in Henderson et al. (1989); z is the caterpillar (unbranched) tree (Fig. 9-6); y the topology with a single branch at the end (Fig. 9-6); x has a single branch in the middle; and topology w has two branches. The numbers for the weighted average distribution are normalized to the number of times each topology occurs (see Fig. 9-6). For any pair of values a and b (e.g., 5 and 3), the columns of the weighted average method will sum to the values for the single column distribution.

In the example in Figure 9-4 [$n = 8$, $a = 5$, $b = 3$ (eight taxa, one color occurring five times and the other three times)] the probability of observing $m = 1$, 2, and 3 changes on the tree is 0.030, 0.303, and 0.667, respectively (Table 9-1). Le Quesne (1989) has derived similar results by direct counting.

Little progress has been made for $r > 2$ colors (codes). If the third and subsequent codes occur only once, then the problem is trivial, for example, $f_m(a,b,1) = (2n-3)f_{m-1}(a,b)$. The single third color can be added to any of the $2n-3$ edges on the trees containing the first two codes. With the third or subsequent colors occurring more than once, the only cases derived are when $m = r-1$ (i.e., with only the least possible number of changes on the tree) and where $m = r$.

$$f_{r-1}(a_1,a_2 \ldots a_r) = (2n-5)!!N(a_1,1) \ldots N(a_r,1)/(2n-2r-1)!! \quad (2)$$

(Carter et al., 1990), and

$$f_r(a_1,a_2 \ldots a_r) = (2n-5)!!N(a_1,1) \ldots N(a_r,1)$$
$$(r-1)(4(n-r)^2-2n+r)/(2n-2r+1)!! \quad (3)$$

(Steel, 1992).

Distributions for all cases up to 16 taxa and four colors have been estimated by simulation.

Equations (1), (2), and (3) are restricted to a single column (character) which limits their usefulness. The reason for this limitation is that the probabilities for each column are not independent. For example, consider a simple two-color (code) case with six taxa where each color occurs three times (*aaabbb*). Figure 9-5 shows the two unrooted topologies for six taxa. (We define topologies as unlabeled trees, rooted or unrooted.) It is not possible to fit such a column to the second topology (Fig. 9-5B) with only a single $a \leftrightarrow b$ change on the tree. If such a column (three *a*'s and three *b*'s) does fit a tree with only one $a \leftrightarrow b$ change, then the tree must be the topology shown in Figure 9-5A. This knowledge will affect the probabilities for subsequent columns of data. Consequently, the above formulae only

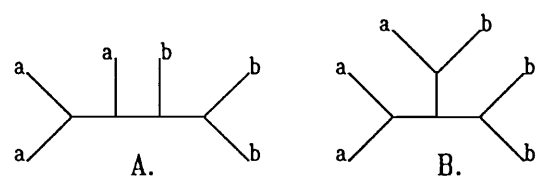


Figure 9-5 Nonindependence of characters. The figure gives the two unrooted topologies (unlabeled trees) for six taxa. A 2-state character with three *a*'s and three *b*'s cannot be fitted to the second topology (B) with only one $a \leftrightarrow b$ change on the tree. If such a column (three *a*'s and three *b*'s) does fit a tree with only one $a \leftrightarrow b$ change, then the tree must be topology (A) whereas four *a*'s and two *b*'s can be fitted to both topologies with a single $a \leftrightarrow b$ change. This indicates how the distribution of changes on different columns of data cannot be independent. The formulae described in the text can only be applied to single columns of data.

apply to single columns of data, although they give good approximations when the data set is large (Steel et al., 1991). We refer to this as the single column distribution.

An alternative approach has been developed by Steel et al. (1991), which takes into account the lack of independence between columns. This evaluates the probabilities for each topology separately. A weighted average is made using the number of trees that can be derived from each topology. The formula for the number of trees from a topology is,

$$n!/2^x6^y \qquad (3)$$

(Hendy et al., 1984),
where x is the number of twofold centers of symmetry in the topology and y (≤ 1) is the number of threefold centers of symmetry (which occurs only with n = 3,6,9, . . .). Figure 9-6 gives the six topologies for n = 9 taxa, identifies the centers of twofold and threefold symmetry, and applies the above formula to each topology. As a check, it is shown that the total number of trees over the six topologies sums to $(2n\text{-}5)!!$ or 135,135. This "weighted average" approach is more difficult to calculate as it requires a separate calculation for each topology, but the results can be combined for many columns of data. It was the method referred to earlier (Henderson et al., 1989), and used for testing nonevolutionary models.

Results of both approaches for up to eight taxa are given in Table 9-1. Weighted average values for n = 9 are given in Henderson et al. (1989). Archie (1989) has recently used simulation to estimate these values.

These methods that calculate the probabilities of finding columns of given lengths on the tree allow some models of evolution that do not assume an evolutionary tree to be tested. In the influenza virus case referred to above, several nonevolutionary models could be eliminated. In the present context, the important point is that a tree is a falsifiable model.

One additional point needs to be considered when calculating the expected number of changes on a tree. Formula (3) assumes that all trees have the same chance of occurring, the "all trees equiprobable" model. (In this context, trees are binary unrooted trees with end-points labeled.) An alternative "Markov model" assumes the trees are derived from a process that includes random speciation and extinction (Simberloff, 1987). Under these circumstances, it is not valid to use formula (3). This second model has been used in biogeographic studies where there are a small number of areas being analyzed. However, the assumptions of the Markov model are not met in many (most?) phylogenetic studies where only a small proportion of possible taxa are used. The subset of taxa in such studies are *not* chosen at random.

In the set of 11 mammals referred to earlier, the taxa used were not randomly selected from all mammalian species. For example, there is only one rodent and no bat sequences. In a random sample of mammalian species, most would be drawn from these two orders. Rather, the taxa have been "selected" to cover a wider range of taxa and so the assumptions

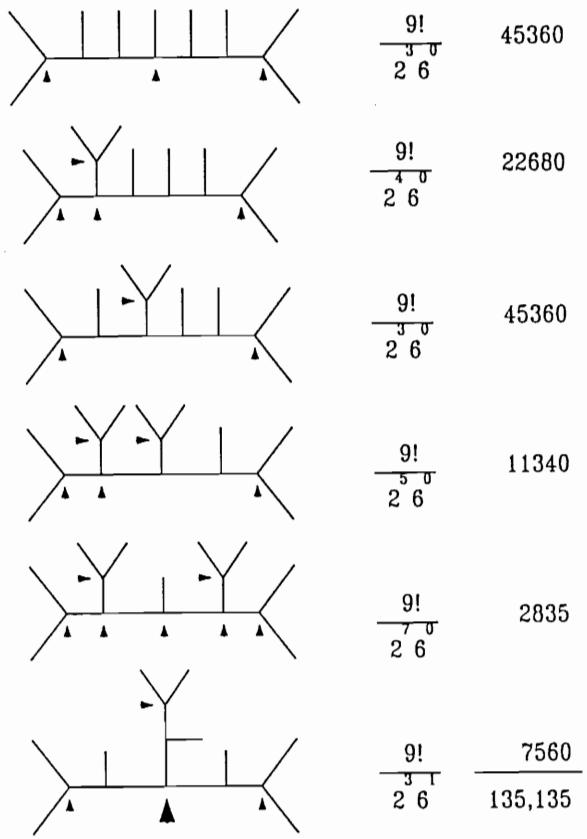


Figure 9-6 Application of the formula for counting trees from topologies. The six unrooted topologies for n = 9 taxa are given. The x centers of twofold symmetry are marked with a small arrow, and the one center of threefold symmetry (y = 1) is marked by a large arrow. The numbers of phylogenetic trees that can be derived from each topology ($n!/2^x6^y$) are shown, together with the observation that these sum to the expected number of trees, $(2n - 5)!!$.

of the Markov model are not met. Under these conditions the neutral "all trees equiprobable" assumption is more appropriate.

As an additional precaution, it should be noted that the probabilities found for the symmetric difference metric vary with the topology. The published values (Hendy et al., 1984, 1988) are a weighted average over all topologies. This is the reason the distribution was initially calculated only for up to 16 taxa where there are about 500 topologies, although more recently some properties have been determined for larger numbers of taxa (Steel, 1988). It is not valid to find one tree and then use the distribution

to estimate the probability that a second tree has x differences. The distribution refers to a pair of trees, selected at random.

VALIDITY OF METHODS FOR INFERRING TREES

Hadamard Transformations

One major conclusion from Popper's approach to science is that more progress is made by trying to find the limits of a hypothesis or conjecture, rather than "testing" the hypothesis in areas where it is expected to apply. In the present context, the hypothesis or conjecture is that a particular tree-building method is expected, given sufficient data, to reconstruct the correct tree. In retrospect, the previous sections are examples of the limitations of making and testing simple predictions. In each case we would have been very surprised if, for example, the trees had not become more similar as longer sequences were used. In our own defense we would say that it is useful to be able to estimate the number of trees that should still be considered possible. Nevertheless, a far more powerful test would have been to try to find conditions under which a method of tree reconstruction would fail.

The work of Felsenstein (1978) and Cavender (1978) has introduced an improved approach which does this. It allows a search for models of evolution where tree building methods would **not** be expected to find the correct tree. That is, it allows a search for conditions where a tree reconstruction method will fail. A method is said to be inconsistent if, under some conditions of the model, it can converge to an incorrect solution as longer sequences are taken. The paradox is that a method may by chance find the correct tree with short sequences, but as longer sequences are used, the probability that the method finds the correct tree goes to 0.

A model of evolution has three parts,

1. a tree (or more generally a graph),
2. an assumed "mechanism" of change to the sequences, and
3. "edge lengths" (probabilities of change along the edges of the tree).

A frequently assumed mechanism (Farris, 1973; Cavender, 1978) is that changes occurring in the sequence are "independent and identically distributed" (i.i.d.). Changes at any position along the sequence, and anywhere on the tree, are independent. All positions (nucleotide or amino acid) have the same chance of changing state (identically distributed). In addition, some mechanisms assume the same rate of change along each edge of the tree—the molecular clock. We will use the term "standard model" for a mechanism of independent and identically distributed changes on a tree but which does not assume the molecular clock. Felsenstein (1978) showed that under some conditions, a model with four taxa and uneven

rates of evolution could give data for which parsimony would be expected to reconstruct the wrong tree. Thus, parsimony is, in general, an inconsistent method even though there are many specific models where it will work correctly. An important question is whether the conditions that lead to inconsistent performance are common (and therefore, parsimony should seldom be used) or unusual (in which case the inconsistency may not be a problem in practice).

To extend Felsenstein's analysis to $n > 4$ taxa we developed the Hadamard transformation (Hendy and Penny, 1989; Hendy, 1991) for 2-state characters, a and b (Table 9-2). We will describe this in three parts, each of which is now straightforward (this was probably not true of the original

Table 9-2 Bipartitions and Vectors for the Hadamard Transformations, Illustrated for the Five Taxa in the Model in Figure 9-8B*

					Probabilities	
Index	Subsets	p	q	δ	r	s
1	{1}	0.100	0.1116	0.1116	0.0000*	0.0583+
2	{1,2}	0.005	0.0050	0.0050	0.2232	0.0113
3	{1,3}			0.0000	0.2332	0.0081
4	{1,2,3}			0.0000	0.2332	0.0155
5	{1,4}			0.0000	0.2332	0.0081
6	{1,2,4}			0.0000	0.2332	0.0155
7	{1,3,4}			0.0000	0.2332	0.0155
8	{1,2,3,4}	0.200	0.2554	0.2554	0.4463*	0.1301+
9	{1,5}			0.0000	0.3720	0.0155
10	{1,2,5}	0.005	0.0050	0.0050	0.3720	0.0113
11	{1,3,5}			0.0000	0.3720	0.0081
12	{1,2,3,5}	0.100	0.1116	0.1116	0.5952*	0.0583+
13	{1,4,5}			0.0000	0.3720	0.0081
14	{1,2,4,5}	0.100	0.1116	0.1116	0.5952*	0.0583+
15	{1,3,4,5}	0.100	0.1116	0.1116	0.5952*	0.0583+
16	{1,2,3,4,5}			−0.7118	0.5952*	0.5197+
				0.0000		1.0000

*The 16 (2^{n-1}) subsets containing 1, of {1,2,3,4,5} for $n = 5$ taxa. In each case the subset with its complement (the remaining taxa) forms a bipartition. Each edge of a tree splits the set of taxa into complementary subsets. Similarly, any 2-state character also splits the set of taxa. We refer to these as "edge bipartitions" and "character bipartitions." The edge bipartitions of the tree of Figure 9-8B have indices 1, 2, 8, 10, 12, 14, and 15. For each edge e_i we list the probability p_i that a character will have different states at its endpoints. The q_i value is the number of changes expected under a Poisson model on that edge. δ is obtained from q by inserting 0 where there is no corresponding q_i value and one negative value for δ_{2^n-1} so that $\Sigma\delta_i = 0$. r is an intermediate vector which contains the actual distance, per nucleotide, of the minimal path between even-sized subsets of taxa. The Hadamard transformation (Hendy and Penny, 1989) allows us to compute s from δ where s_j is the probability of obtaining the character bipartition with index j. Thus, for c characters, cs_j will be the expected frequency of this character bipartition. Note in this case $s_{16} = 0.5197$, so we would expect approximately 52% of characters to be constant, while bipartition 3 should occur for less than 1% (0.81%) of characters. The inverse Hadamard transformation produces a vector δ, with $q_i = \delta_i > 0$ identifying the edges e_i of the tree.

Parsimony uses only "informative" characters, those which group taxa together (that is, ignoring singleton and constant characters). Their corresponding s_i values are marked "+." In choosing the parsimony tree(s) for these data we find four minimal length trees (expected length 0.0902c), none of which is the tree used to generate the data. The expected length of the correct tree is 0.0934c, the third longest of the 15 possible binary trees. This is an example of inconsistency in parsimony, even with equal rates of evolution.

description). The three parts are the model, the expected form of the data, and a method for interconverting between these two (that is, calculating the data from parameters of the model, and *vice versa*.

We note that each edge e_i of a tree partitions the taxa into two subsets, the set of taxa to the left of the edge and the set of taxa to the right. This pair of sets is called an edge bipartition (split). There are 2^{n-1} possible bipartitions. Also, each column of the sequences induces a bipartition of the taxa, the two subsets being those taxa with state *a* and those with *b*. These we call sequence bipartitions.

We express the bipartitions by the taxa (represented by numbers) they contain. For example, if the character-states for four taxa are *a*, *b*, *b* and *b* (summarized as *abbb*), this is the bipartition {1} and {2, 3, 4}. (The character-states *baaa* also give this bipartition.) There is no need to list both subsets of the bipartition. We normally list only the subset containing taxon 1.

There are four (2^{n-2}) ways in which the sequence bipartition {1} {2, 3, 4} can be generated on a tree and these are shown in Figure 9-7. They correspond to the four ways of labeling two ($n-2$) internal points with the two codes. The probabilities for each of the eight (2^{n-1}) bipartitions can be calculated in this way. In all there are 2^{2n-3} sets of calculations (2^{n-2} ways for 2^{n-1} partitions). However, the whole procedure can be carried out using Hadamard transformations. Assume a tree T where for each edge e_i there is a probability p_i that the character-states at each end of the edge are different. With these values given we use the Hadamard transformation to calculate the probability of obtaining any particular sequence bipartition.

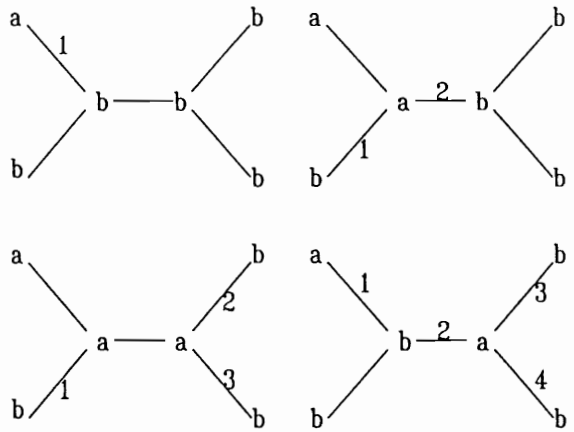


Figure 9-7 Ways of forming a single bipartition for four taxa. The figure shows four combinations of changes along edges of a tree that all lead to the bipartition {1} {2,3,4}.

The transformation is

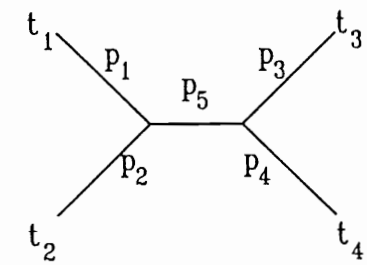
$$\mathbf{s} = (\mathbf{H}'\ln(\mathbf{H}\,\boldsymbol{\delta}))/2^{n-1} \qquad (4)$$

where $\mathbf{H}$ is a Hadamard matrix of 2^{n-1} rows and columns, and $\mathbf{s}$ and $\boldsymbol{\delta}$ are defined in Table 9-2. Equation (4) is easily inverted as $\mathbf{H}^{-1} = \mathbf{H}'/2^{n-1}$ giving

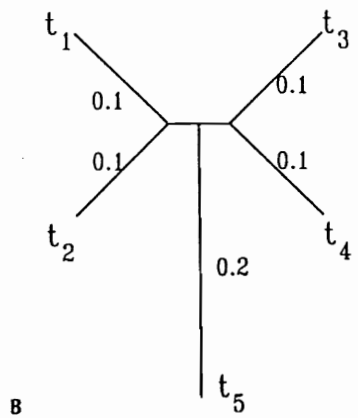
$$\boldsymbol{\delta} = \mathbf{H}\exp((\mathbf{H}'\,\mathbf{s})/2^{n-1}). \qquad (5)$$

The calculations are illustrated for a case with $n = 5$ in Table 9-2 with an example derived from Figure 9-8B. In practice it is not necessary to construct the Hadamard matrix, as there is now a simple algorithm to carry it out. From equation (4) we can find values of edge lengths on a tree where parsimony will be inconsistent; that is, it will be guaranteed to select the wrong tree as sequences become longer.

$\boldsymbol{\delta}$ and $\mathbf{s}$ are equivalent for a specified mechanism of change in that each can be obtained from the other, and are not dependent on a particular tree. Calculating the expected data in this way quickly allows the tree to be identified to which a tree-building method will converge. Then the rate



A



B

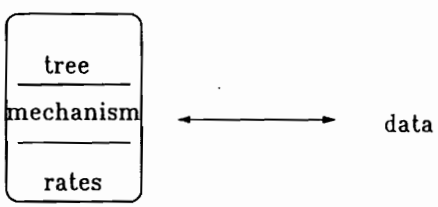
Figure 9-8 Trees used in calculations. (A) A tree of four taxa (t_{1-4}) with p_i being the probability, for any column, of observing a change along an edge. p_i can take values between 0 (no change) and 0.5 (complete randomization). $\omega_i = (1 - 2p_i)$. Conditions under which several tree-building methods will converge to the correct tree are given in the text. (B) The unrooted tree used to illustrate the Hadamard transformations (see Table 9-2). The edge length of the central edges are 0.01. The tree is consistent with the molecular clock if rooted on the central pendant edge.

of convergence to this tree can be studied separately by randomly selecting larger data sets from s.

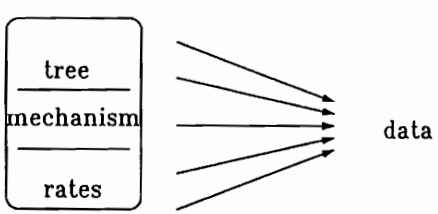
The computational advantage of the Hadamard transformations over existing maximum likelihood methods is that they allow the data (sequences or bipartitions) to be used directly for estimation of parameters of the model. This is indicated in Figure 9-9a to show that it is possible to go from the model to give predicted sequence bipartitions, or from sequence bipartitions to estimate lengths of edges on trees.

By contrast, existing maximum likelihood methods start with a single tree and some initial guesses for edge lengths. It then repeatedly (maybe thousands of times) makes slight adjustments to edge lengths and repeats the calculation to see if the observed data are now more likely. The process is indicated in a general way in Figure 9-9b to indicate the repeated calculations. Figure 9-9b has been modified from the usual calculation for maximum likelihood to allow an easier comparison to the Hadamard transformations. One interesting development would be to combine the Hadamard transformations with maximum likelihood as the optimality criterion. Such a process would set a good estimate of the optimal branch lengths with a single computation.

The optimality criterion we have favored is finding the closest tree (Hendy, 1991). Its advantages come from its being characterized mathematically.



A Hadamard transformations



B maximum likelihood

Figure 9-9 The advantage in being able to invert the calculation. With the Hadamard transformations each calculation need only be calculated once, not repeatedly as for maximum likelihood.

This allows the global optimal tree to be found exactly (by branch and bound methods). Thus far, the closest tree method is the only one based on the idea of general invariants for $n > 4$. We have used the closest tree criterion to find the optimal tree for 20 taxa (Penny et al., 1990). For 20 taxa there are $> 10^{20}$ trees! By contrast, current maximum likelihood methods would not search more than 100 (10^2) trees—an advantage of 10^{18} times for methods based on the Hadamard transformations.

Conditions for Consistency—Four Taxa

Some progress has been made with four-taxa, 2-state codes (colors) on the general conditions under which a method will, or will not, converge to an incorrect tree on the standard model. Using the terminology of Figure 9-8A and defining $\omega_i = (1 - 2p_i)$ we find some common tree-building methods *will* converge to the correct tree (be consistent) if and only if, the following conditions are met (Steel, 1989).

Cluster Analysis. For a simple clustering procedure which joins the pair of taxa having the smallest observed distance, the condition for consistency is:

$$\omega_5 < \frac{\max\{\omega_1\omega_2, \omega_3\omega_4\}}{\max\{\omega_1\omega_3, \omega_1\omega_4, \omega_2\omega_3, \omega_2\omega_4\}}.$$

It can be shown that this is a special case of the next condition shown below. Thus, for example, if parsimony fails to converge to the correct tree with four taxa, so too will a simple clustering procedure. The converse does not hold. The special case, with $p_1 = p_3$ and $p_2 = p_4 = p_5$ was derived by Felsenstein (1978).

Parsimony or Compatibility Methods and Methods Using the Four-Point Distance Criterion. This last method selects $\min\{d_{1,2} + d_{3,4}, d_{1,3} + d_{2,4}, d_{1,4} + d_{2,3}\}$. These are consistent, if and only if,

$$\omega_5 < \min\left\{\frac{(\omega_1\omega_2 + \omega_3\omega_4)}{(\omega_1\omega_3 + \omega_2\omega_4)}, \frac{(\omega_1\omega_2 + \omega_3\omega_4)}{(\omega_1\omega_4 + \omega_2\omega_3)}\right\}.$$

With equal rates of evolution (the molecular clock) this condition is always met and parsimony will be consistent with four taxa.

Maximum Likelihood, Closest Tree (from Hadamard Transformations), and the Corrected Four-Point Distances Metric. These methods will always be consistent with four taxa on the standard model. The corrected four-point condition is to select,

$$\min\,(d_{ij} + d_{kl} - 2 \cdot d_{ij}\, d_{kl}/c) : \{i, j, k, l\} = \{1, 2, 3, 4\}$$

where c is the number of columns.

It should be noted that, at least with parsimony, the conditions for consistency are more restrictive as the number of taxa increases. Parsimony with:

four taxa, can only be inconsistent with unequal rates of evolution;
five taxa, with equal rates of evolution (the molecular clock) can be inconsistent if the root is on the central pendant edge which is adjacent to a short and long edges;
six taxa, with arbitrarily low rates of change can be inconsistent;
n large, with all branch lengths (internal and pendant) are both small and equal can be inconsistent.

The first three cases (four to six taxa) are from Hendy and Penny (1989) and the last is from Steel (1989). The conclusion is that the range of cases where parismony is inconsistent increases as more taxa are added.

It is desirable that this type of study be extended to other optimality criteria. With four taxa, methods using the four-point distance criterion have identical performance (with respect to consistency) as parismony (see above). Again with four taxa, simple clustering methods are worse than parsimony in that they will fail under a wider range of conditions. We know less about the performance of cluster analysis with additional taxa except that it is not identical to parsimony. With five or more taxa we can find examples where parsimony fails and simple clustering is consistent, and *vice versa*.

Loss of Information

We have commented elsewhere (Penny et al., 1990) that all methods ignore some of the information in sequences. With four taxa, distances use nearly all the information, but the proportion of information in distances declines very rapidly (Steel et al., 1988) as the number of taxa increases.

Table 9-2 also indicates information that is omitted by parsimony and standard distance methods. Singletons and constant partitions (marked "+" in Table 9-2) are omitted by parsimony methods. Entries in **r** which do *not* correspond to a path between a pair of taxa (marked "*") are omitted by distance methods. There is no general correspondence between these omissions in **r** and **s**. The losses of information in parsimony and distances are not equivalent. Parsimony methods ignore any singletons and constant columns, corresponding to $n + 1$ elements in **s**. Distance methods use all the **s** values to construct a distance matrix which corresponds to only $n(n - 1)/2$ of the **r** values. The Hadamard transformations use all of these 2^{n-1} values. Table 9-3 contrasts the increase in the numbers of bipartitions used by the Hadamard transformation with those used by parsimony and distance methods for small values of n. The example in Table 9-2 and Figure 9-8B is a case where the loss of information is sufficient for parsimony to converge to the incorrect tree (Hendy and Penny, 1989). The correct tree found by selecting partitions {1, 2} and {1, 2, 5} is not the shortest, but the third longest tree (out of 15).

Table 9-3 Comparative Use of Information by Parsimony and Distance Methods*

	Bipartitions		
Taxa	Total	Used in Parsimony	Entries in Distance Matrix
4	8	3	6
5	16	10	10
6	32	25	15
7	64	56	21
8	128	119	28
9	256	246	36
10	512	501	45
11	1024	1012	55
12	2048	2035	66
13	4096	4082	78
14	8192	8177	91
15	16384	16368	105
16	32768	32751	120

*For n = 4–16 we list the number of bipartitions, together with the numbers considered by parsimony and distance matrix methods. Parsimony methods omit $n + 1$ bipartitions from **s** which rapidly become a minute proportion of the total. The omitted bipartitions include the n singletons and the number of constant columns. These omitted bipartitions are important under any model that includes estimates of rates of change. Because they occur so often, the bipartitions omitted by parsimony are those with the most accurate estimates of their frequency. Distance methods include some information from all bipartitions, but the proportion of entries of **r** that are used, rapidly becomes very small (Steel et al., 1988).

Twelve Mammals, Seven Sequences

In earlier papers, we used a data set from 11 mammals and five (later extended to six) sequences. The data are amino acid sequences converted to "best-guess" nucleotide sequences. These data have been particularly useful for developing and testing new methods of analysis. However, even six sets of sequences were insufficient for the optimal tree to be stable (that it would not change as longer sequences became available). When the minimal and near-minimal trees were analyzed (Penny and Hendy, 1985b) it was found that the position of the carnivore (dog) was the least stable. The carnivore was attached by a long "unbranched edge" which can be (Hendy and Penny, 1989) comparatively unstable on the tree. Would adding a second carnivore improve the stability? Would we then get a tree that would be expected to remain constant as longer sequences became available? In other words, would we have convergence to a single tree?

We have now added a second carnivore to the tree and an additional sequence, α-crystallin, for all 12 taxa. Thus the data set now has seven sets of sequences from 12 mammals (Table 9-4). The sequences from the earlier studies were α- and β-hemoglobins, fibrinopeptides A and B, cytochrome *c*, and myoglobin. In some cases a "taxon" was made up of sequences from two species. One example is using either mouse or rat to represent the rodent taxon. The additional carnivore taxon is made up

ould have
o Figures
ne correct
nates).
s, relative
ition with
n the tree
after the
ungulate
ed from a
next two

hat would
e ungulate
We again
dge, and a
fficient to
o possibil-
tree, they
give incor-

correction
. However
edges of a
columns.
rsimony to
pated. The
he problem
with equal
tween long
r or not in-

lusion. The
illion years
divergence,
ence) to be
billion, or
ected to be
ry fortunate
(rRNA)] is
out testing.
nowledge of
is sufficient.
can only be
onstruct the
ee from One


```
                    1234567890123456789012345678901234567890123456789012345678901234567890
                    GACGGGGGGAUUACGCAAAAAACCCGGGGCCUCUGCAGAGCGGCCAGGCACGCCCCGAUCACGCGGGCGAGACCAAACAAAGGCUUADAACGAGCAGAC
1. Monkey           CU.....G..CU.GGAAA....G.....CAA.A....C...U.C.....G.....C...C.A...G.....C.....
2. Sheep            .G.....C.....GGC.....G....A.A..G.AACAGA..AA.....U.....G....A....U..A....G.G.A..G.G.C...UCAU.C.
3. Horse            .....G.....CU.G.....A.GU.AAC....U..GU...UU....A.C...AC.....C.UGAGC.C...GC...G.G..C...C..C..
4. Kangaroo         .GG.A.....CA..A...GC..AA...AAA.G...A..GC.....AA..AG.AAACCA.A.A..C..U.G.A...A....AAAA.....A.GC..
5. Rodent           .U..A...A.....C....G..G..C.....A..C.....AU....U...G.....U....A..GAC.....C...UC.AAA..C.C..C..U...
6. Rabbit           C..A.AA.....GAA.....G.....A..AA.U....A.....G.AU.....A.....U..A.AA..C..A.GA.U.....A.....A...G.....C.....C.
7. Dog              C..AAAA..CCA.....C.....U.GAA...GCU.....U.U..G...AC.....C..A.GA.U.....A.....A...G.....C....AU.
8. Pig              .G.....C....GA..G.....G.A.....AACA.A.....U...UU.G.....A.....U.A.....AG.G.CCA.UC...GU
9. Cat              .U..U.A...CG.....C..C...A.AC...U.GA....GCUU.U....G.A.G.....U...AA.GA.U.....A.G.....G..CCAA...U.GU
10. Human           C.....C.....G.....C.GAAA....G.....CU.....C.....G.....GCC.....C....G.....CC.....U.
11. Cow             .GG.....C.....GGC.....A.....A....AGC..A..AA.....GUA...CA.....AA...U..AC...GGGC...G.GCC...UCAU.C.
12. Ape             C.....C.C.....G.....C.GAAA....G.....CU.....C.....G....GCC.....CA...G.....CC.....U.
                                110       120       130       140       150       160       170       180       190
                    1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901
                    GGGAGGAACCGAGCCGAGAGGAUUGAGCAGCUUACACGAGCGUGAGCAGGUACCAAACCCGAGCAAUGGACCGGAAACAAGAAADCGCC
1. Monkey           .....C.....AA....A.CC.AGAAUU.A.C.CUA.....C.....GU.....U.....U....G.A.C..A.UA.
2. Sheep            .U.G.....CA.CUCC.....A.C.G..C..AC....GAGA.C..CCGA.G....C.U.555555555555
3. Horse            .A.....A.....ACA..U.C.....A....C..A.....C..AC...CC...AC..CC...C..A.....G....C.A..G
4. Kangaroo         AA...AC.A.....UA.A..GG.....A.UC.....A.AG.G.....UC....C.....C.A.....UC...AA.A
5. Rodent           ..ACA...AAC...GCA..U.A...AG....G....C.....C.G.AA...G.....A.....G..A.....G..U.G.....
6. Rabbit           ....AAC.AAC...GC.....UU.C.....G....C.A.....C....AC.....CC..CG..A.C...G.A..G.....
7. Dog              AA.....A....A..AU.UA..CG.....CA.A.AC...AAA.GG..C....C.....C.A.A.....G.U.GC.G...
8. Pig              CAA.....CCA..UCAG.....A...G.....C.....G.....G...GC.G...
9. Cat              .....GCA.....A.....U..G.G.....CA...AC....A...GG..C...A...CCG.C.A.A....A.G.U.GC.G...
10. Human           .....C...UA.AA.....U..AGAAUU.A.CUCUA.....C.....GC.....UG.....A.U.C..G.A.C..A.UAA
11. Cow             CCAG...C.....CCA..UGG..G.....A.C.G..C..AC....GAGAAC..CCGA.G..U...UA.GA...C.G...
12. Ape             .....C...UA.AA.....U..AGAAUU.A.CUCUA.....C.....GC.....UG.....A.U.C..66666666666
```

*Columns 1–42 and 43–97 are derived from α- and β-hemoglobins; 98–110 and 111–127 from fibrinopeptides A and B; 128–142 from cytochrome *c*; 143–179 from myoglobins; and 180–191 from α-crystallin. Columns 1–142 are (with the addition of cat) from Penny et al. (1982), columns 143–179 from Penny and Hendy (1986), and columns 180–191 are derived from deJong et al. (1977). α-crystallin sequences were not available for sheep or apes and so additional singleton character-states ('5 and 6') were assigned to these taxa. This does not distinguish between any tree, and so does not affect the results obtained. The amino acid sequence for lion cytochrome *c* was supplied by Professor R.P. Ambler, Edinburgh. The data, with some additional information, are available by e-mail from D.Penny @ massey.ac.nz.

Figure 9-10 Trees for 12 mammals, based on seven sequences. (a) is the minimum length tree (472 mutations) based on the 191 columns in Table 9-4. It is also the consensus or median tree (Penny et al., 1982). (b) is three changes longer (475 mutations) but may be correct if the "long edges attract" problem of parsimony is leading to incorrect convergence. One additional change results from transferring carnivores and two changes from separating the rabbit-rodent neighboring pair.

from six feline sequences (five from cat and lion cytochrome *c*) plus one from a seal. The procedure of using composite taxa is legitimate in this study as it cannot result in trees from different sequences being more similar than expected. It can only introduce more dissimilarity between trees if the composite taxon was phylogenetically incorrect. For the reasons discussed below we suggest the best tree is that shown in Figure 9-10b.

Another approach is possible by analyzing the frequency of the edges in minimal trees from bootstrap samples. This is illustrated in Figure 9-11 with results from 132 minimal trees (in this case using parsimony) from bootstrap samples. Only the relationships between three groups; primates, lagomorphs (rabbit), and rodents are shown. The most common subtree is (rabbit, rodent), primate (Fig. 9-11b) which occurs 74 times. The next most frequent is (primate, rabbit), rodent (Fig. 9-11a), 44 times; and (primate, rodent), rabbit subtree (Fig. 9-11c), which occurs seven times. Figure 9-11b conforms to the Glires model (Novacek, 1985) which links rodents and lagomorphs (rabbits).

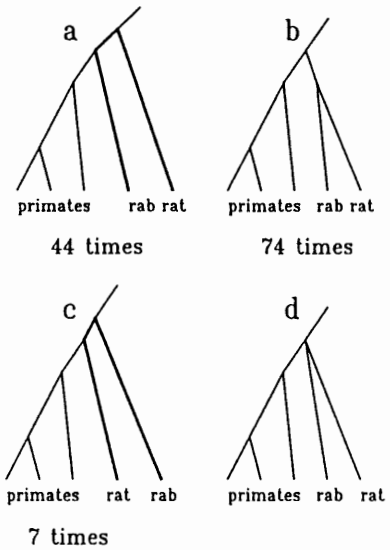


Figure 9-11 Frequency of partitions in bootstrap and jackknife samples. (a), (b), and (c) are the three subtrees for primates, rabbit (rab), and rodents (rat), and the frequency of occurrence of these subtrees in the minimal length trees from 132 bootstrap samples. (d) is the unresolved trichotomy. An anomaly is the low frequency of just one of the three subtrees, (c) compared to (a). This can be explained as an example of the error with parsimony involving a path of long-short-long edges, as in the heavy lines in (a). With parsimony these long edges would tend to "attract" (Hendy and Penny, 1989) leading to tree (b).

The question is to interpret these frequencies (74, 44, 7). If there was a short edge linking these three groups (that is, close to a trichotomy, Fig. 9-11d) then all three subtrees (Figs. 9-11a–c) would be expected to occur with about equal frequency. If Figure 9-11b was correct, then we would expect Figures 9-11a and 9-11c with equal frequency. The observed frequencies are significantly different from those expected for either the trichotomy (Fig. 9-11d) or Glires (Fig. 9-11b) models.

Qualitatively we can account for the observed distribution of Figure 9-11a–c if Figure 9-11a ((primate, rabbit), rodent) is the correct relationship. This tree has a long-short-long series of edges (rabbit, internal edge, rodent) which is shown as dark lines in Figure 9-11a. Parsimony would lead to the long edges being drawn together to give Figure 9-11b. This would account for two of the three subtrees being found more frequently than the third. A prediction from this hypothesis is that bootstrap samples from data sets which include sequences from a quite different lagomorph (or rodent) should resolve the issue. Note that convergence from the other two trees (Fig. 9-11b and 9-11c) could not lead to the observed frequencies in that neither should lead to the observed frequencies. If Figure 9-11b

were correct, then any tendency for incorrect convergence should have equal affects on both rodents and lagomorphs, leading equally to Figures 9-11a and 9-11c. We suggest the tree in Figure 9-10b shows the correct relationship of the three groups (rodents, lagomorphs, and primates).

A similar problem occurs with the position of the two carnivores, relative to the ungulates and the metatherians (kangaroo). The bipartition with the two carnivores is almost always found, but their position on the tree varies. It is on the first division of the eutherians (83 times), after the separation of ungulates (40 times), or on the first division of the ungulate lineage (eight times). Again we do not have the pattern expected from a near-trichotomy (all three equal, or one more common and the next two equal). The difference is statistically significant in a χ^2 test.

Is this another case of incorrect convergence? Consider what would happen if the correct tree had the carnivores separating after the ungulate line has diverged (that is, in the position shown in Fig. 9-10b). We again have a case of a long lineage (kangaroo), a short internal edge, and a medium lineage leading to the two carnivores. This may be sufficient to indicate incorrect convergence. Again, if either of the other two possibilities for kangaroo, ungulates, and carnivores was the correct tree, they would not have the long-short-long sequence of edges that can give incorrect convergence.

These conclusions need to be made quantitative. A possible correction for parsimony is available (Penny and Hendy, in preparation). However it requires a reasonable estimate of rates of change along the edges of a tree. This depends on the numbers of singletons and constant columns.

We note that the "long edges attract" problem that leads parsimony to converge to a wrong tree may be more frequent than anticipated. The initial work by Felsenstein (1978) with four taxa suggested the problem may occur with very unequal rates. With five taxa it can occur with equal rates (Hendy and Penny, 1989) but required a short edge(s) between long edges. Even this knowledge is inadequate to indicate whether or not incorrect convergence occurs in a specific case.

Even with these data we were unable to reach a firm conclusion. The time of separation of these 12 mammals is probably 60 to 80 million years ago. If seven sequences are insufficient to resolve such a recent divergence, would we expect a single sequence (even quite a long sequence) to be sufficient for a divergence that occurred much earlier—say 1 billion, or even 3 billion years ago? Such a rhetorical question is expected to be answered in the negative. It is of course possible that we are very fortunate and a single sequence [e.g., small-subunit ribosomal RNA (rRNA)] is sufficient. But we have no rational reason to *assume* it without testing. Work on the small-subunit rRNA has vastly improved our knowledge of the phylogeny of organisms but there is no reason to assume it is sufficient.

What concerns us is the scientific myth (and at present, it can only be considered a myth) that a single sequence is sufficient to reconstruct the whole history of life. We call this the "Myth of a Universal Tree from One

Gene" (MUTOG). The myth is strong enough to get evolutionary trees into textbooks which lack qualifying statements. The main problem is that the myth inhibits testing of ideas by not taking additional sequences and testing for convergence. If the tree is "believed" to be correct, there is little motivation to undertake new work to test the tree. Such a state of affairs is disturbing to anyone using a Popperian framework for science. An idea (trees in this case) should be a stimulus to new measurements and better, more rigorous tests.

DISCUSSION

By discussing our work from a Popperian viewpoint, we do not wish to imply a concept of a single monolithic framework in which scientists work. Evolutionists in particular should be well aware of the problem of essentialism. They are aware of the importance of diversity in evolution (Mayr, 1982), and consequently should be skeptical of any attempt to define "one true scientific method." Medawar, who supported a Popperian view of science (Medawar, 1974), has commented that

> Among scientists are collectors, classifiers and compulsive tidiers-up. Many are detectives by temperament and many are explorers; some are artists and others artisans. There are poet-scientists and philosopher-scientists and even a few mystics (Medawar, 1967:132).

Attempts by philosophers to describe a single mechanism of research that all successful scientists follow, are doomed to failure. A diversity of approaches to science is still accommodated on the Popperian model in which no scientific hypothesis is ever absolutely proven, where no hypothesis should be "believed." A similar comment on the diversity of approaches, using the analogy of diversity within a species appears in Hull (1988). The important point is the attitude toward hypotheses: they should never be accepted as beyond questioning. They are tools to aid in the design of harder and more rigorous tests.

We have recently discussed tree-building methods as requiring at least five criteria. Methods should be consistent, robust, efficient, fast, and falsifiable (CREFF) (Penny et al., 1990). These criteria appear incompatible in that efficient methods (which increase in a polynomial manner with the number of taxa) for real data are not known. Little is known on the robustness of methods relative to deviations from the assumptions about the mechanism of change. Improvements are still needed to increase the power of tree-building methods. Perhaps the most urgent is an analog of the Hadamard transformations for four-state characters. So far, an initial approach has been developed (Penny et al., 1990) which still needs further development. Another area that is important to test is the effect of deviations from the "standard model." This would allow a test of the robustness of tree building.

Whatever the approach that is taken, it appears to us that it is important to find the limits of any tree-building method and to find under what

conditions it will break down. Such a Popperian approach should help identify problem areas and assist the search for better methods. Improved methods will almost certainly depend on a better understanding of their mathematical basis. This makes the study of evolutionary trees an exciting part of modern biology.

REFERENCES

Allan, M. (1977) *Darwin and His Flowers: The Key to Natural Selection*. Faber and Faber, London.

Archie, J.W. (1989) A randomization test for phylogenetic information in systematic data. *Syst. Zool.* **38**:239–252.

Carter, M., M.D. Hendy, D. Penny, L.A. Székely, and N.C. Wormald. (1990) On the distribution of lengths of evolutionary trees. *SIAM J. Disc. Math.* **3**:38–47.

Cavender, J. (1978) Taxonomy with confidence. *Math. Biosci.* **40**:271–280.

Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Day, W.H.E. (1985) Optimal algorithms for comparing trees with labelled leaves. *J. Classif.* **2**:7–28.

Day, W.H.E., D.S. Johnson, and D. Sankoff. (1986) The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.* **81**:33–42.

de Jong, W.W., J.T. Gleaves, and D. Boulter. (1977) Evolutionary changes of α-crystallin and the phylogeny of mammalian orders. *J. Mol. Evol.* **10**:123–135.

Farris, J.S. (1973). A probability model for inferring evolutionary trees. *Syst. Zool.* **22**:250–256.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401–410.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.

Ghiselin, M.T. (1969) *The Triumph of the Darwinian Method*. University of Chicago Press, Chicago.

Graham, R.L., and L.R. Foulds. (1982) Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosci.* **60**:133–142.

Henderson, I.M., M.D. Hendy, and D. Penny. (1989) Influenza viruses, comets, and the science of evolutionary trees. *J. Theor. Biol.* **140**:289–303.

Hendy, M.D. (1991). A combinatorial description of the closest tree algorithm for finding evolutionary trees. *Disc. Math.* (in press).

Hendy, M.D., C.H.C. Little, and D. Penny. (1984) Comparing trees with pendant vertices labelled. *SIAM J. Appl. Math.* **44**:1054–1067.

Hendy, M.D., and D. Penny. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* **59**:277–290.

Hendy, M.D., and D. Penny. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**:297–309.

Hendy, M.D., M.A. Steel, D. Penny, and I.M. Henderson. (1988) Families of trees and consensus. Pp. 355–362 in *Classification and Related Methods of*

Data Analysis (H.H. Bock, ed.). Elsevier Science Publishers. B.V., Amsterdam.

Hoyle, F. (1984) *Living Comets*. Cardiff University Press, Cardiff.

Hull, D. (1988) *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, Chicago.

Le Quesne, W.J. (1989) Frequency distribution of lengths of possible networks from a data matrix. *Cladistics* **5**:395–407.

Mayr, E. (1982) *The Growth of Biological Thought: Diversity, Evolution and Inheritance*. Harvard University Press, Cambridge.

Medawar, P.A. (1967). *The Art of the Soluble*. Methuen, London.

Medawar, P.A. (1974) Hypothesis and imagination. Pp. 241–273 in *The Philosophy of Karl Popper*, Vol. 1 (P.A. Schlipp, ed.). Open Court, LaSalle.

Mickevich, M.F. (1978) Taxonomic congruence. *Syst. Zool.* **27**:143–158.

Novacek, M. (1985) Cranial evidence for rodent affinities. Pp. 59–81 in *Evolutionary Relationships Among Rodents: A Multidisciplinary Analysis* (W.P. Luckett and J.-L. Hartenberger, eds.). Plenum Press, New York.

Penny, D. (1985) The evolution of meiosis and sexual reproduction. Biol. J. Linn. Soc. **25**:209–220.

Penny, D., L.R. Foulds, and M.D. Hendy. (1982) Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297**:197–200.

Penny, D., and M.D. Hendy. (1985a) Testing methods of evolutionary tree construction. *Cladistics* **1**:266–278.

Penny, D., and M.D. Hendy. (1985b) The use of tree comparison metrics. *Syst. Zool.* **34**:75–82.

Penny, D., and M.D. Hendy. (1986) Estimating the reliability of evolutionary trees. *Mol. Biol. Evol.* **3**:403–417.

Penny, D., M.D. Hendy, E.A. Zimmer, and R.K. Hamby. (1990) Trees from sequences: panacea or Pandora's box? *Aust. Syst. Bot.* **3**:21–38.

Popper, K.R. (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, London.

Popper, K.R. (1972) *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, Oxford.

Popper, K.R. (1976) *Unended Quest: An Intellectual Autobiography*. Fontana, London.

Popper, K.R. (1978) Natural selection and the emergence of mind. *Dialectica* **32**:339–355.

Popper, K.R. (1984) Erkenntnis und gestaltung der wirklichkeit: die suche nach einer besseren welt. Pp. 11–40 in *Vorträge und Aufsätze Aus Dreissig Jahren*. R. Piper, Munich. English translation: (1991) Knowledge and the shaping of reality: the search for a better world. New Zealand Sci. Rev. 48 (in press).

Riddiford, A., and D. Penny. (1984) The scientific status of evolutionary theory. Pp. 1–38 in *Evolutionary Theory: Paths Into the Future* (J.W. Pollard, ed.). John Wiley and Sons, London.

Robinson, D.F., and L.R. Foulds. (1981) Comparison of phylogenetic trees. *Math. Biosci.* **53**:131–147.

Ruse, M. (1975) Darwin's debt to philosophy: an examination of the influence of the philosophical ideas of John F.W. Herschel and William Whewell on the development of Charles Darwin's theory of evolution. *Stud. Hist. Philos. Sci.* **6**:159–181.

Schweber, S.S. (1977) The origin of the *Origin* revisited. *J. Hist. Biol.* **10**:229–316.

Simberloff, D.S. (1987) Calculating probabilities that cladograms match: a method of biogeographical inference. *Syst. Zool.* **36**:175–195.

Steel, M.A. (1988) Distribution of the symmetric difference metric on phylogenetic trees. *SIAM J. Disc. Math.* **1**:541–551.

Steel, M.A. (1989) *Distributions on bicoloured evolutionary trees*. Ph.D. Dissertation, Massey University, Palmerston North.

Steel, M.A. (1992) Distributions on bicoloured binary trees arising from the principal of parsimony. *Discr. Appl. Math.* (in press).

Steel, M.A., M.D. Hendy, and D. Penny. (1988) Loss of information in genetic distances. *Nature* **336**:118.

Steel, M.A., M.D. Hendy, and D. Penny. (1991) Significance of the length of the shortest tree. *J. Classif.* (in press)

Thompson, E.A. (1975) *Human Evolutionary Trees*. Cambridge University Press, Cambridge.