6.7 Have we got the true tree?

Given the range of possible methods for inferring phylogeny, we naturally want to know if they work, that is, do they recover the actual evolutionary relationships among nucleotide sequences? Several approaches have been developed to answer this question: analysis, simulation, known phylogenies, and congruence.

6.7.1 Analysis

In some cases, the phylogenetic method is simple enough that we can establish mathematically the exact conditions under which that method would fail. A good example of this is UPGMA, which requires a molecular clock. This condition can be expressed more formally in terms of branch lengths for a three-taxon tree (Fig. 6.26). This amounts to requiring that the two most closely related taxa are more similar to each other than they are to any other taxa. If this is not the case, then using UPGMA will generate an erroneous tree (Fig. 6.27).

Parsimony methods have been much debated and much studied. Early on it was shown by Felsenstein that under a very simple model of evolution with

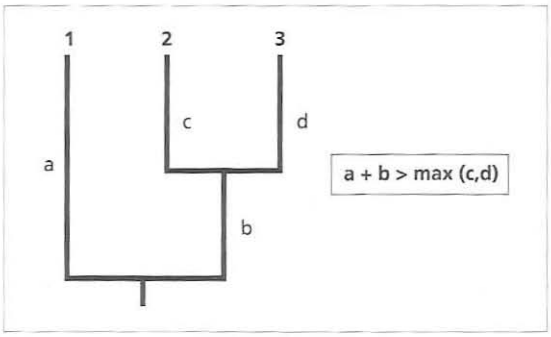


Fig. 6.26 The condition required for UPGMA to successfully reconstruct the true tree is that the sum of the edges leading to sequence 1 (a + b) must be greater than the larger of c and d. After Mooers *et al.* (1994).

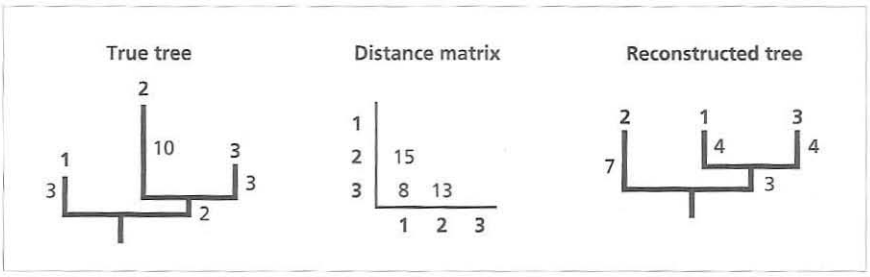


Fig. 6.27 An example where UPGMA will reconstruct the wrong tree. The edge lengths on the true tree violate the condition shown in Fig. 6.26, as a + b = 3 + 2 < max (c, d) = 10. Sequence 2 has evolved more rapidly than the other two sequences, so that sequence 1 and 3 are more similar to each other than either is to 2.

just two character states, parsimony could be inconsistent (see section 6.4.4). Originally it was thought that this inconsistency was due to the unequal rates of evolution in Felsenstein's example, hence parsimony might be consistent if rates of evolution were more equal. However, counter-examples have been found for five or more sequences where parsimony is inconsistent even if rates of evolution are constant. These examples serve to emphasise that it is not unequal rates of evolution *per se*, but the distribution of edge lengths that can cause problems for parsimony (see also Box 6.3).

While analytical methods are elegant and can yield important insights into phylogenetic methods, for more than a few sequences and for more complicated models of evolution, analytical methods become increasingly intractable. This has prompted the use of other strategies, such as experimental phylogenies and simulation.

6.7.2 Known phylogenies

The most compelling evidence for the success of our tree-building methods would be that they could reconstruct known phylogenies, that is, phylogenies which we knew to be true from other evidence. Unfortunately, known phylogenies are very rare. Typically the only 'known' phylogenies are those of laboratory animals and crop plants, and even these are often suspect. David Hillis and colleagues have addressed this problem by creating 'experimental' phylogenies in the laboratory. By subdividing cultures of bacteriophage T7 grown in the presence of a mutagen they constructed a known phylogeny (Fig. 6.28). At the end of the experiment, they obtained restriction site maps and nucleotide sequences for the eight T7 cultures. This data was then input into various tree-building programs, the output of which was compared with the actual T7 tree.

For the restriction site data, UPGMA, neighbour-joining, Fitch & Margoliash, Cavalli & Sforza and parsimony all recovered the topology of the actual tree. Because they had the actual ancestor cultures, Hillis *et al.* were also able to compare those ancestors with the reconstructions obtained by parsimony. Fully 97.3% of the ancestral states were correctly inferred.

Although encouraging, this particular study is perhaps less informative than

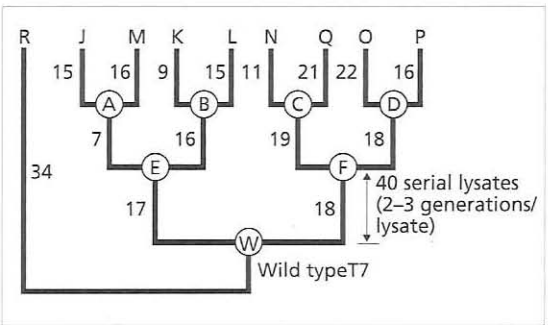


Fig. 6.28 Artificial phylogeny for bacteriophage T7 constructed by culturing the phage in the presence of a mutagen. Letters in circles are ancestors, numbers on branches are numbers of restriction site differences between phages at each node. After Hillis *et al.* (1992).

might have been hoped. The tree is well balanced and all nodes are accompanied by numerous changes making the tree a relatively easy one to reconstruct; indeed, every method tried was successful. More will be gained by fabricating more difficult trees and discovering under what circumstances different methods fail.

6.7.3 Simulation

Very rarely do we have known evolutionary trees for actual organisms or sequences, whether obtained from nature or 'grown' in a laboratory. An alternative source of known trees is computer simulation. In this case, we supply a computer program with a tree and 'evolve' DNA sequences along the branches of the tree according to some model. We then provide the resulting sequences as data for a range of tree-building methods, and determine which tree-building methods succeed in recovering the original tree. An advantage of this approach is that we can explore the effects of a wide range of parameters on the performance of tree-reconstruction methods. A disadvantage is that the models used to generate the artificial sequences may be unrealistic, making it difficult to generalise the results of the simulations to actual data sets (although see 'parametric bootstrapping' in section 6.8.3). In particular, we need to avoid biasing the model towards a particular method. For example, if we used the model implicit in one tree-building method to generate the sequences, then we would expect that method should perform very well, perhaps better than any other method. However, this result does not allow us to claim that this method is therefore the best tree-building method—all we can say is that under the chosen model the method works well.

A group of researchers lead by David Hillis has used simulation to explore the relative merits of a range of tree-building methods for the simplest phylogenetic problem—finding the unrooted tree for four sequences. They varied the overall rate of nucleotide substitution, and the rate of substitution in different parts of the tree. These two parameters can be represented by a diagram showing the simulation space (Fig. 6.29).

By generating many thousands of data sets for different combinations of branch lengths and recording the success of various methods in recovering the actual trees upon which the data was generated, a picture emerges of the overall performance of the different methods.

Figure 6.30 shows the relative performance of UPGMA and parsimony under the same conditions. As we would expect, UPGMA does best when the rate of evolution is relatively constant, hence the high degree of success along the diagonal line in the chart. However, the further a tree departs from this diagonal line the less successful UPGMA is in recovering it. Parsimony does not require a molecular clock, and this is reflected in the wider range of branch lengths that parsimony can accommodate. However, there is a region in the top left of the parameter space, corresponding to a short internal edge and two long terminal edges. In this region, often called the 'Felsenstein zone', is where

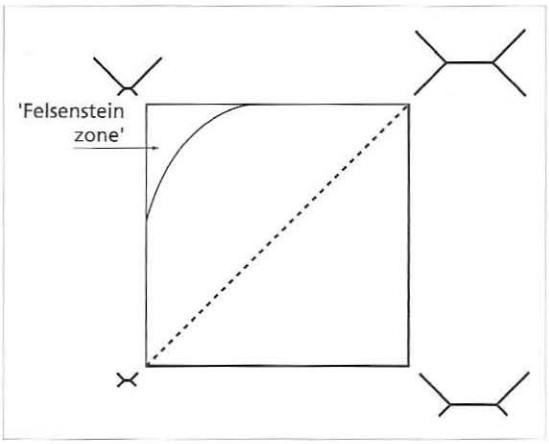


Fig. 6.29 The tree space explored by the computer simulations analysed in Fig. 6.30. At each corner of the square the corresponding tree is drawn. Along the dotted line the rate of substitution is the same on each edge in the tree. The trees at top left and bottom right show two extremes in rate variation, differing in how much change occurs along the central edge. The region at the top left-hand corner has become known as the 'Felsenstein zone', and corresponds to parameters Felsenstein (1978) used to show that parsimony could be inconsistent. After Huelsenbeck and Hillis (1993: Fig. 4).

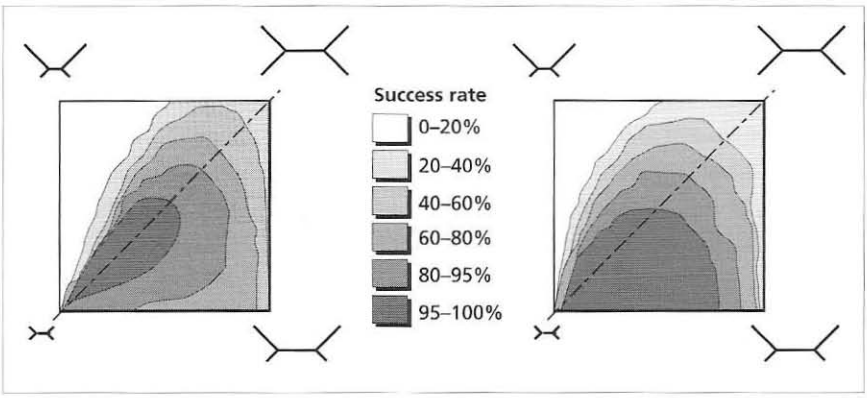


Fig. 6.30 Performance of UPGMA (left) and parsimony (right) in recovering the true tree under the range of parameters shown in Fig. 6.29. The success rate is the percentage of times that the correct tree was recovered in that region of the parameter space. Note the white region in the top left of the two diagrams (the 'Felsenstein zone') where neither method performs well. After Huelsenbeck and Hillis (1993).

'long branches attract' (see section 6.4.4), misleading parsimony into inferring the wrong tree.

Another variable affecting the performance of phylogenetic methods is the amount of data available. By creating artificial data sets of different sizes, it is

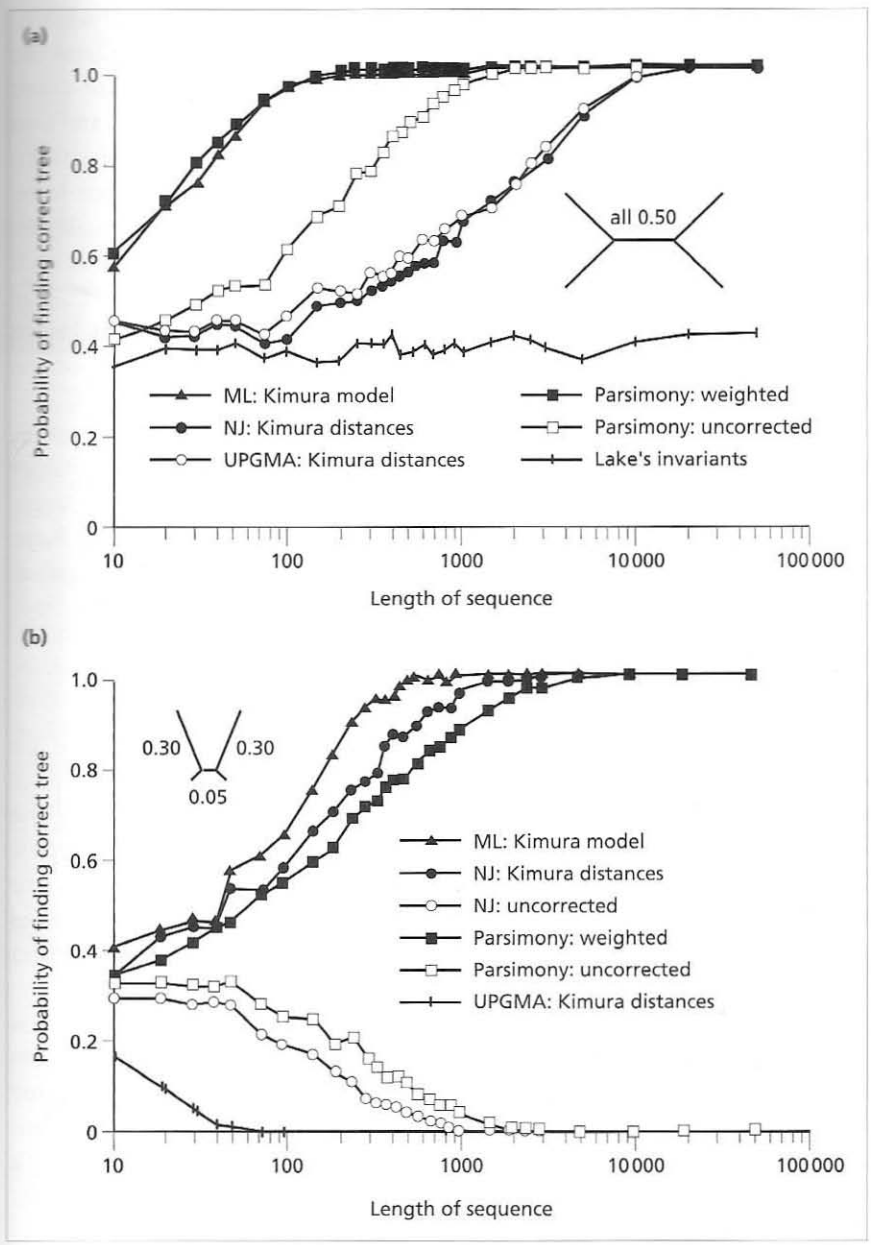


Fig. 6.31 The accuracy of several different phylogenetic methods in reconstructing two four-taxon trees with (a) all edges equal in length and (b) with a short internal edge and two long terminal edges. In each graph the proportion of analyses that recovered the correct tree is plotted against the length of the simulated sequences. From Huelsenbeck *et al.* (1996: Figs 2 and 3).

possible to see how if adding more data improves a method's rate of success. Figure 6.31 shows the results of two simulation experiments to evaluate the accuracy of several different phylogenetic methods in reconstructing four-taxon trees with all edges equal in length and with a short internal edge and two long terminal edges. In the case of equal branches all the methods are consistent, but differ in their efficiency (Lake's method, which is not discussed here, requires about 10 million base pairs to converge!). The tree with unequal rates of evolution requires more data to accurately reconstruct, and methods such as unweighted parsimony, neighbour joining on uncorrected distances, and UPGMA all converge on the wrong answer and are hence inconsistent. Hence the accuracy of different tree-building methods can vary depending on the tree being reconstructed.

The four-taxon tree studied by Hillis and colleagues is the simplest possible, and may be unrepresentative of the problems faced when we try and infer larger phylogenies (see Box 6.3). The difficulty of simulation experiments on larger trees is that as the trees get larger the numbers of combinations of tree shape and edge lengths also increases, prohibiting the kind of thorough investigation possible for the four-taxon case.

6.7.4 Congruence

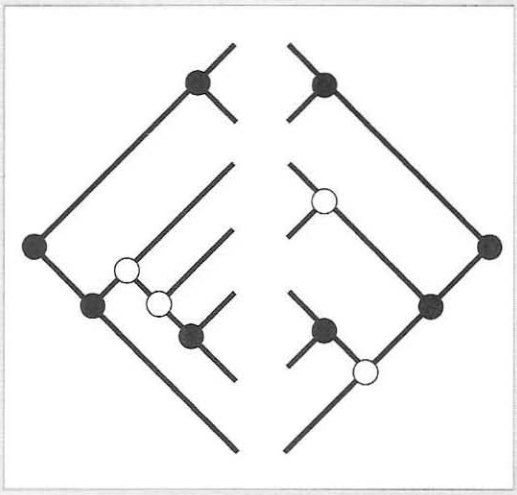
Congruence is the agreement between estimates of phylogeny based on different characters. If the data sets are independent then the probability of obtaining the same or even similar trees for the same organisms using the different data sets is vanishingly small for any reasonable number of species. This is a simple consequence of the large numbers of possible phylogenies (see Chapter 2). Hence, if different data sets give us similar trees this gives us confidence that both reflect the same underlying cause, namely they reflect the same evolutionary history.

The power of the congruence test stems from the assumption that there is a single phylogeny and that each data set being analysed should reflect that history, coupled with the improbability of obtaining similar phylogenies due to chance alone (Box 6.7). Congruence has been used in two related, but distinct ways: validating a method of phylogenetic inference, and validating a particular source of data.

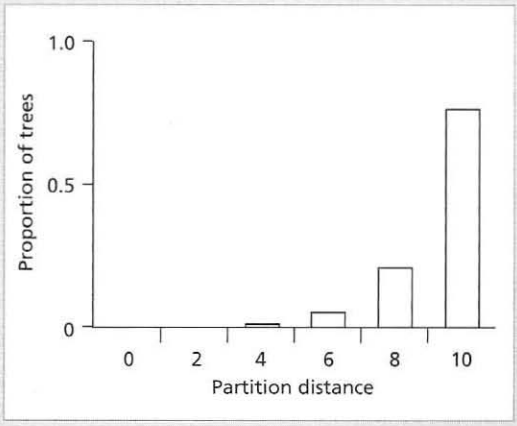
Validating a method of inference

If two data sets for the same set of taxa contain phylogenetic information then we might expect the trees inferred from those two data sets to be similar, if not the same. This follows from the assumption that each data set has the same evolutionary history. Hence, a method that consistently recovered similar trees from different data sets would be preferred to a method that produced different trees from different data sets. An early illustration of this approach using DNA

Box 6.7 How similar are two trees?



In this example the two trees each share four identical splits (solid circles) and each have two unique splits (open circles), making a total of four unique splits. Hence the partition distance between the two trees is four. As each tree has five internal nodes, two maximally dissimilar trees would have a partition distance of 10. The chart below shows the distribution of this measure of tree dissimilarity if pairs of trees are drawn at random from the set of all possible trees for seven sequences. The bulk of trees are as different as it is possible to be, and fewer than 0.6% of trees share as much in common as the two trees shown above.



sequences is the study by Penny and colleagues. Using parsimony they found minimal and near-minimal length trees for five different proteins (cytochrome c, fibrinopeptide A and B, haemoglobin α and β) and asked whether the trees

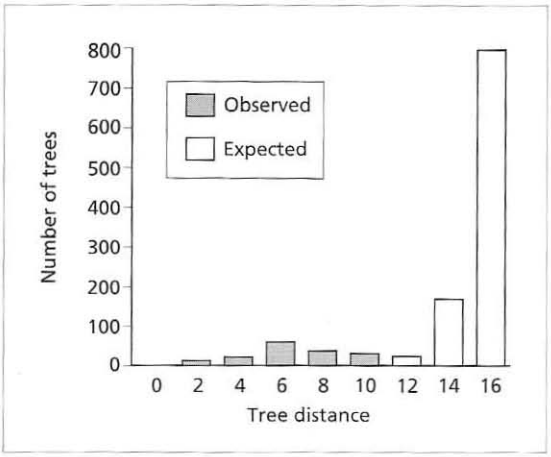


Fig. 6.32 Observed distribution of pairwise distances between 39 parsimony trees for 11 mammals inferred from five genes and the expected distribution if the trees were sampled at random. Note that the trees from the five genes are much more similar than could be expected due to chance alone.

were more similar to each other than could be expected due to chance alone using the partition distance (Box 6.7). The trees parsimony analysis obtained from the different genes were much more similar than could be expected due to chance alone (Fig. 6.32).

Validating a new source of data

As an example of the second use of congruence, we could ask whether a newly sequenced gene, or a new measure of genetic variation, contained phylogenetic information by comparing trees constructed using those data with trees from other data sources. Indeed, this test can be applied to any potential source of phylogenetic information. Figure 6.33 shows trees for modern humans based on genetic data and languages. The trees are different, but more similar than could be expected due to chance alone, suggesting that some common cause underlies this similarity. The obvious possibility is that both the genes and the language of modern human populations reflects the historical relationships among those populations.

6.8 Putting confidence limits on phylogenies

In the previous section we discussed accuracy, that is, how close we are to the truth. Now we turn to precision. In molecular phylogenetics our goal is to recover 'the one true tree' for a set of sequences. A measure of the phylogenetic precision is how many trees we can reject as candidates for the true tree. If we reject all but one, then we have maximum precision. Of course, this sole surviving tree may not be the true tree for our method might have misled us (remember that precision is not the same as accuracy; see Box 6.6, p. 208). The greater the number of trees we cannot reject, the

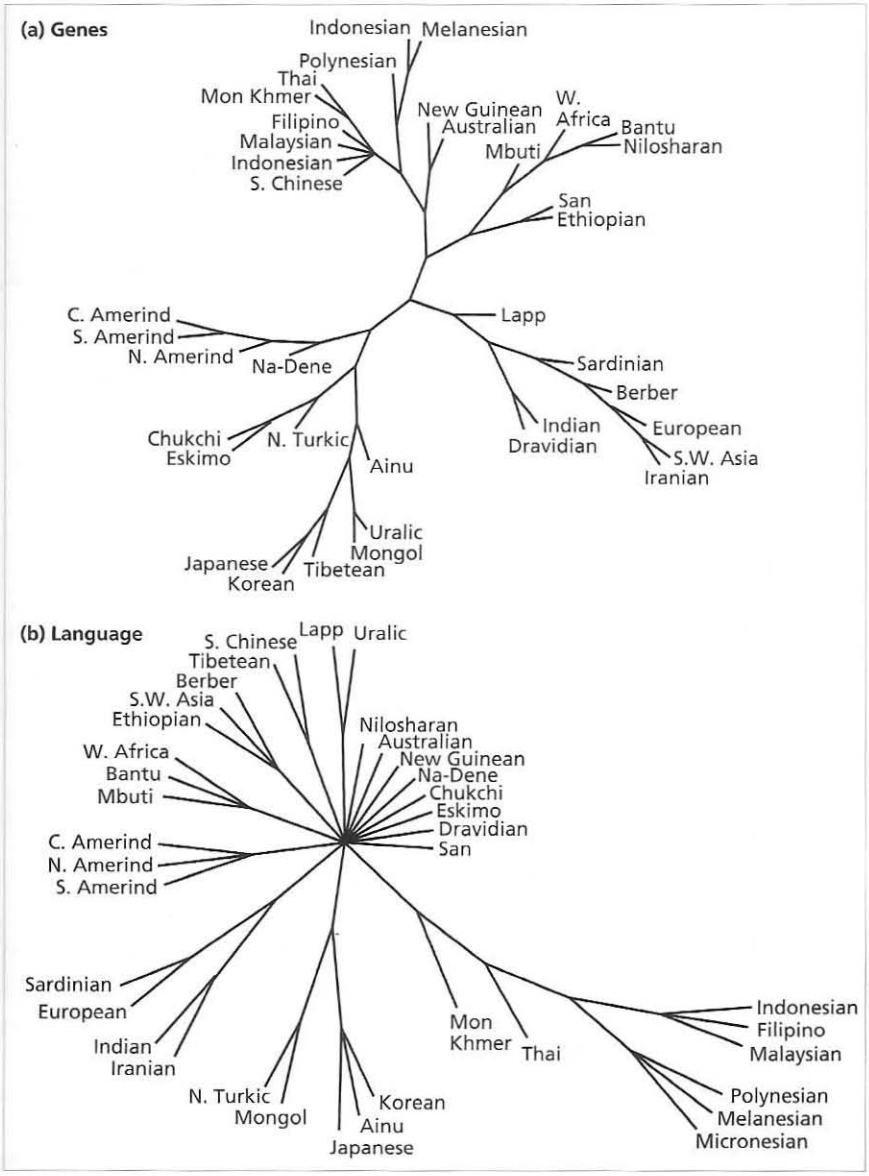
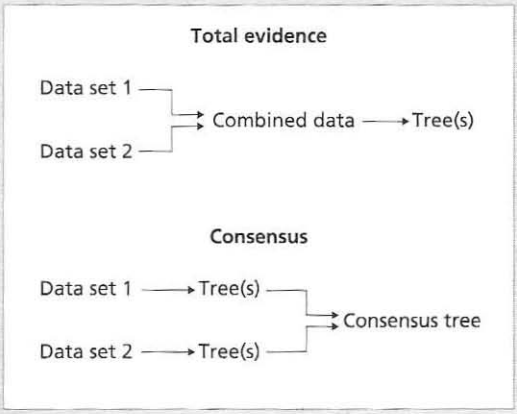


Fig. 6.33 Evolutionary trees for modern humans derived from genes and from language characteristics. The two trees are different but more similar than two randomly chosen trees. This suggests that both genes and language reflect the history of modern humans. After Penny *et al.* (1993).

lower the precision. Most scientific measurements are accompanied by some estimate of precision, such as 13.5 ± 0.2 mm. Phylogenetics should be no different, hence estimates of phylogeny should be accompanied by some indication of confidence limits.

Box 6.8 Congruence and total evidence

Suppose that the trees shown in Box. 6.7 are for the same organisms but obtained from different genes. The trees show points of similarity and points of difference. If our goal is to obtain a tree for the organisms themselves we have a dilemma—what is the best estimate of the phylogeny of these organisms? Two different solutions are to combine the two data sets and analyse them as one data set ('total evidence') or to analyse the data separately and combine the resulting trees for each data set using a consensus tree (see Chapter 2).



There are arguments for and against both views. Combining data sets makes use of all the data, whereas combining trees can result in less than optimal solutions. However, analysing data sets separately can reveal whether the different data sets support similar trees. If they do not then it may be that the genes have quite different histories and the data sets should not be combined to estimate organismal phylogeny (Bull, 1993) (see Chapter 8 for the gene tree/species tree problem).

6.8.1 Sampling error

One reason for a poor estimate of a phylogeny may not be the method used, but the data itself. If a data set contains homoplasy then different nucleotide sites support different trees, hence which tree (or trees) a given data set supports will depend on which characters have been sampled. As a consequence, estimates of phylogeny based on samples will be accompanied by sampling error.

As an example, consider the 896 base pairs of mitochondrial DNA used by Brown and colleagues to study hominoid phylogeny. Of these 896 sites, 90 are phylogenetically informative using parsimony (Fig. 6.34). The most parsimonious tree for all 90 sites is ((human, (chimp, gorilla)), orang, gibbon). However, consider what would happen if we sampled only the first 31 base pairs, recovering just five informative sites. The first such site supports (human,

```
Human      GTCATCATCCTTCTTTTTTTAGCAATTTCCTCACCTTCTCCGTCACGCTC 50
Chimpanzee A.T.C...T.C.T....CCCC...T.C...CTG......T.A.T.T.TCT 50
Gorilla    ..TG-T..TACCTCCC...C.A...CCC.T.TGTT.CAC.TA..G..TC. 50
Orang-utan AC..CTCC.ACC...CC.CCTAAG.C.CA.A...TCAACT..C...A.CT 50
Gibbon     AC.GC.CC.A.C.CC.CCC.CAAGTCC.ATC..T.CAA..TACTGTA..T 50

Human      TCGCCGCTCTCACTCCCCTTATTTTCTTGTCCGGTGACCG 90
Chimpanzee C.....T..C..T.TT...C........ACT.A..... 90
Gorilla    C..T.AT..CA...TT.......C.T.C.C.TA....TTA 90
Orang-utan CTATTA.CT.AGTC..TACCGCC.AGCCA.TTCACACTAA 90
Gibbon     .TA.TA.CT.AG.C..TACAGCCCAGCCAAA..ACACTAA 90
```

Fig. 6.34 The 90 phylogenetically informative sites from the Brown *et al.* (1982) mitochondrial DNA data for hominoids.

gorilla), the second site supports (human, chimp, gorilla), the third site contradicts the first and supports (chimp, gorilla), site four groups gorilla and orang, and site five again supports human + gorilla. The most parsimonious tree for this subset is (((human, gorilla), chimp), orang, gibbon), which is not the most parsimonious tree for the complete data set. This example is rather contrived, but it illustrates the point that results may depend on the sample of sequences. Indeed, the 896 sites sequenced by Brown *et al.* represent in themselves only a small fraction of the approximately 16 000 base pairs that comprise the mammalian mitochondrial genome, and more extensive mtDNA data sets support the tree (((human, chimp), gorilla), orang, gibbon).

The effects of sampling error can be seen by comparing trees for different mitochondrial genes. The mitochondrial genome is inherited without recombination, and so every gene has the same phylogenetic history. Árnason and colleagues recently compared trees for different mitochondrial genes from mammals for which the complete mtDNA sequence was known. While most genes agreed on the same tree, others supported different trees (Fig. 6.35).

Just as we often have limited samples of DNA, we may also have limited samples of taxa. In the example just given, mammalian relationships were being inferred based on just six species, a tiny fraction of the total extant Mammalia. Relationships among sequences can change if additional sequences are added. Indeed, for clades with a good fossil record, morphological data may prove superior to molecular data because the evidence from the extinct taxa is crucial to recovering the actual phylogeny.

6.8.2 Estimating sampling error: the bootstrap

One way to measure sampling error is to take multiple samples from the population being studied and compare the estimates obtained from the different samples. The spread of those estimates gives us an indication of the extent of sampling error, that is, how much our conclusions would vary depending on

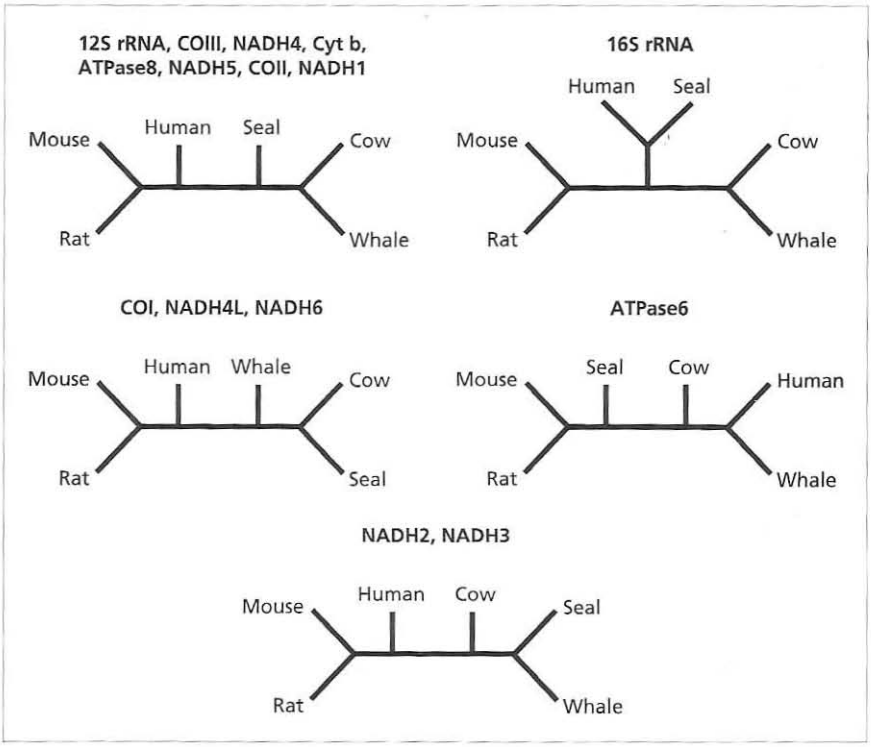


Fig. 6.35 Phylogenies for the same six mammals based on 15 different mitochondrial genes. After Årnason and Johnsson (1992).

what sample we took. Given that repeated sampling is often expensive, it would be nice to be able to calculate sampling error without having to take multiple samples. For many simple distributions (for example, the normal distribution) there are simple equations for calculating confidence intervals around an estimate (for example, the standard error of a mean). Trees are rather complicated structures, and it is extremely difficult to develop equations for confidence intervals around a phylogeny. Hence the reliance on other means of measuring confidence intervals, such as the computer intensive methods of jack-knifing and bootstrapping. Of these two the bootstrap is most often used in phylogenetics.

In a sense the bootstrap mimics the first method of estimating sampling error, but instead of sampling from the population it resamples from our sample. Each resampling is a **pseudoreplicate**. From each pseudoreplicate we derive an estimate of the parameter we are trying to measure, such as the mean height of a population. The variation among estimates derived from each pseudoreplicate is a measure of the sampling error associated with the parameter. For example, we could use the standard deviation of the means of the pseudoreplicates as an estimate of the standard error of our original estimate of the mean (Fig. 6.36).

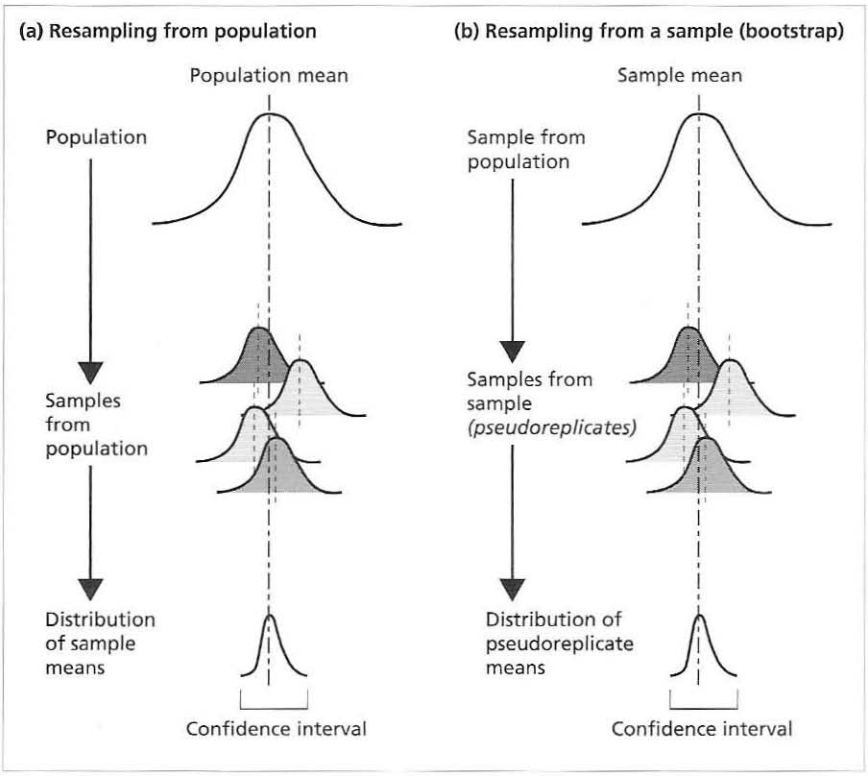


Fig. 6.36 Two methods for estimating the sampling error of the estimate of population mean. One method (left) is to take repeated samples from the population and obtain the distribution of the mean of each sample. An alternative method is to apply the same method but to a single sample (right). Instead of resampling from the original population, the single sample is itself repeatedly sampled to generate pseudoreplicates, and the distribution of the mean of each pseudoreplicate is used to estimate the sampling error of the original estimate of the population mean.

Bootstrapping can be applied to phylogenies by generating pseudoreplicates from the sequence data. For example, given the 896 nucleotide hominoid data set, we could generate a single pseudoreplicate by sampling at random and with replacement from the original data set until we had a new data set comprising 896 nucleotide sites. Because we are sampling with replacement (i.e. any site sampled is 'returned' to the data set before the next sample is taken) some sites may occur more than once in the pseudoreplicate, while others may not be represented at all. Hence the pseudoreplicate will resemble the original data set in that it contains only sites found in that data set, but it will differ in the frequencies of different sites. From this pseudoreplicate we would then build a tree using any of the methods described in this chapter. We then repeat this two-step process a large number of times (anywhere from 100- to 1000-fold), resulting in a set of bootstrap trees. This set of trees contains information on the sampling error associated with our sample.

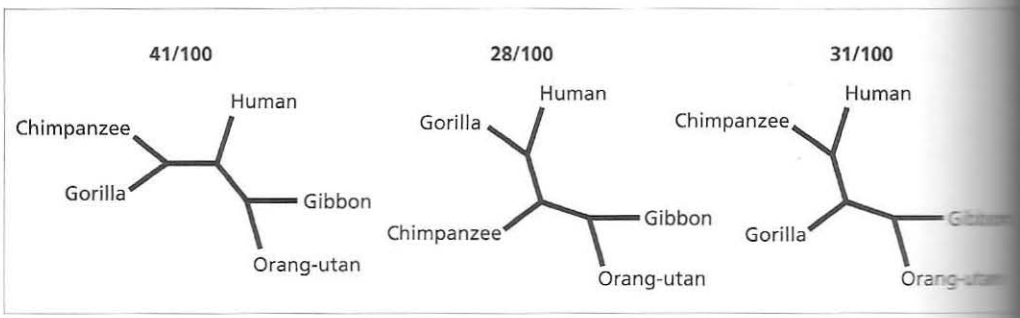


Fig. 6.37 The three trees for the hominoid sequence data that are obtained from 100 bootstrap pseudoreplicates and their relative frequencies. All three trees have the split {{orang, gibbon} {human, chimp, gorilla}} but they disagree about relationships between humans and the African apes.

Bootstrapping the hominoid data 100-fold produces 100 trees, each of which is one of the phylogenies shown in Fig. 6.37. The split {orang, gibbon} {human, chimp, gorilla} occurs in all the trees and hence has a bootstrap value of 100%. However, there is clearly conflict about relationships among the humans and African apes, with all three possible hypotheses receiving almost equal support. This result suggests that, on the basis of this data set, we lack sufficient information to discriminate between these three trees, and that different samples may favour different resolutions of human–chimp–gorilla relationships purely due to accidents of sampling.

For small numbers of sequences, such as the five hominoid mtDNA sequences, it is feasible to show the frequencies of the bootstrap trees. However, for larger numbers of sequences this becomes impractical. Instead, the most common splits found among the bootstrap trees can be assembled into a bootstrap consensus tree (see Chapter 2). These are often drawn with each node labelled with its frequency of occurrence among the bootstrap trees. When originally applying the bootstrap to phylogenies, Felsenstein suggested that only nodes with bootstrap values above 95% should be accepted as well supported. It is important to stress that bootstrap values estimate precision, not accuracy (see Box 6.6). A node may have a high bootstrap value but be completely wrong. In particular, if a tree-building method infers the wrong tree for a given data set, bootstrapping that data set may yield a robust, but wrong answer.

6.8.3 Parametric bootstrapping

'Parametric bootstrapping' resembles a cross between the bootstrap and simulation. Like simulations, it involves generating artificial data using a computer. However, whereas the simulations discussed in section 6.7.3 were conducted using made up trees and simple evolutionary models, in parametric

Box 6.9 Caveats concerning the bootstrap

One important caveat concerning the bootstrap is that this technique makes the assumption that nucleotide sites are independent and identically distributed (i.d.d.). This means that each site is independent of every other site, and that there is a single distribution of rate of evolutionary change across all the sites. In other words, if we were sampling from the mitochondrial genome this assumption means we could treat the genome as a single, homogeneous, entity. Cummins *et al.* (1995) made a simple test of this hypothesis by comparing the tree for 10 complete mitochondrial genomes with trees computed from two types of samples from those genomes: contiguous blocks of sequences, and randomly chosen sites scattered throughout the genome. The first method mimics how the mtDNA genome is sampled in practice; typically a single gene, or part of that gene is sequenced. If the i.i.d. assumption is valid then both kinds of samples should yield equally good approximations to the tree for the whole genome — in fact the randomly scattered sites performed better. Hence, the i.i.d. assumption is not valid for mtDNA. It is not known how robust the phylogenetic bootstrap is to violation of this assumption.

A further consideration is that the results of bootstrapping are often summarised using a majority-rule consensus tree (consensus trees are discussed in Chapter 2) showing the frequency of each split that occurs in at least half of bootstrap trees (splits that are compatible with these splits but occur in less than half the trees may be subsequently added to this consensus tree). If one or more sequences have uncertain relationships they may appear in very different positions in the bootstrap trees (i.e. they 'float' over the tree), resulting in a general lowering of bootstrap values for those parts of the tree over which the sequences float. Hence, parts of the tree which are actually quite robust may have spuriously low bootstrap values. This problem can be addressed using other consensus methods (Page, 1996a; Wilkinson, 1996).

bootstrapping the evolution of the DNA sequences is simulated using parameters (including the tree) estimated from real data. For example, analyses of 18S rRNA sequences suggests that birds and mammals are sister taxa (each other's closest living relative), whereas morphological evidence from both living and fossil taxa groups birds with crocodiles. One possible reason for this conflict between molecules and morphology is that the 18S rRNA data has suffered from long branch attraction (section 6.4.4). Certainly the tree resulting from parsimony analysis of these data shows long edges leading to both the birds and mammals (tree 1, Fig. 6.38), but this is not of itself evidence for long branch attraction. What we want to know is even if tree 3 was the real tree for the 18S rRNA data (as suggested by morphology) could our analyses mistakenly conclude that the bird–mammal tree (tree 1) was best for these data? If this was the case then this would be evidence for long branch attraction.

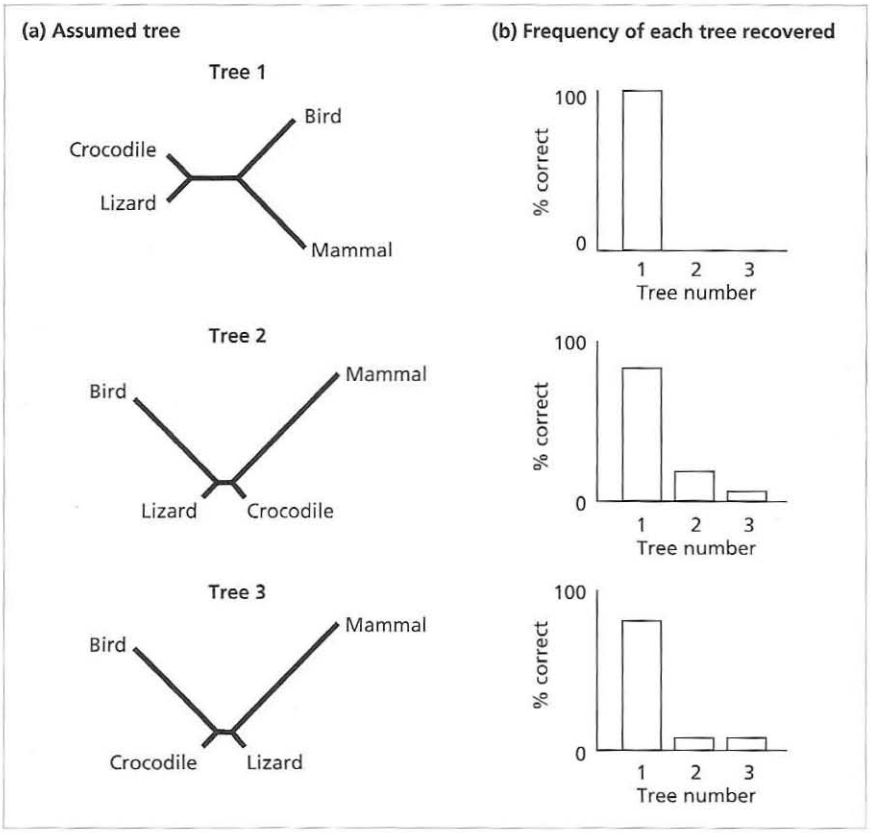


Fig. 6.38 Parametric bootstrapping. Three alternative trees for 18S rRNA sequences from a bird, mammal, crocodile and lizard are shown on the left. For each tree 1000 artificial data sets of the same length as the original 18S rRNA data were generated using parameters derived from that tree. On the right is shown the proportion of times each tree was the most parsimonious tree for the data sets derived from each tree. Note that no matter which tree was used to generate the data, tree 1 is most often recovered as the most parsimonious tree. After Huelsenbeck *et al.* (1996).

This question can be investigated by simulating the evolution of 18S rRNA on the three possible trees for the four taxa and seeing how often a tree-building method recovered the correct tree. Huelsenbeck *et al.* (1996) performed this analysis by first estimating the branch lengths, shape parameter of the gamma distribution (see Chapter 5), and transition/transversion ratio for the real 18S rRNA data set on each of the possible trees for the four taxa. They then used these parameters to generate 1000 artificial data sets of the same size as the original 18S rRNA data set. The results (Fig. 6.38) show that no matter which tree the sequences were evolved on, tree 1 was recovered at least 85% of the time. Even if tree 3 is the true tree (as suggested by morphology), the 18S rRNA data would support tree 1. This suggests that the 18S rRNA data may indeed have been affected by long branch attraction.

6.8.4 What can go wrong?

There are several sources of error that may confound our attempts to reconstruct the true evolutionary relationships among a set of sequences. Many of these have been covered above, but we will summarise them here.

Sampling error

Almost all phylogenetic inference is based on samples of characters, hence the conclusions obtained will in part reflect the vagaries of sampling. This is especially true given the ubiquitous presence of homoplasy, hence the importance of trying to gauge the confidence limits around a tree. Although our focus has been on sequence relationship as opposed to organismal phylogeny, the latter is one of the major uses of molecular phylogenies. In this context a further sampling problem arises, namely the requirement that species relationships are based only on orthologous sequences. If this requirement is not met, organismal and gene trees can become confounded. This topic is considered further in Chapter 8.

Incorrect model of sequence evolution

Methods of phylogenetic inference make assumptions, either implicit or explicit, about the evolutionary process. If these assumptions are violated then we may be misled. Basic assumptions made include variation of rate of substitution among sites, among branches in the tree, and among different nucleotides. To give but one example, the frequencies of the four nucleotides can vary considerably among organisms. If, for example, a gene is AT-rich then most substitutions are likely to involve these two nucleotides simply due to chance. If another, unrelated organism, is similarly AT-rich then it is quite likely that they will share similarities due to independent parallel substitutions, and consequently may be incorrectly grouped together due to this similarity in base composition.

Tree structure

In some cases evolutionary history itself may conspire to thwart our efforts to recover it. Rapid successive cladogenesis, widely differing rates of divergence and extinction can all compromise our ability to accurately reconstruct evolutionary trees. Both simulations and analytical studies have shown that there are situations when current tree-building methods perform poorly. Despite our best efforts, the trace of history may have been eroded by subsequent evolution.

6.9 Summary

Different methods may be more appropriate in different circumstances. This conclusion may seem disturbingly inconclusive, but it reflects the nature of the problem. Maximum likelihood is a very attractive method but it is computationally very expensive, which limits its usefulness as an optimality criterion for trees. Its strength lies in its explicit statistical basis, making it a powerful tool for analysing models of evolution. For all but relatively small data sets, maximum likelihood's main use is likely to be investigating molecular evolution on given phylogenies, rather than as a tool for finding those phylogenies. While advocates of maximum likelihood often look askance at methods such as parsimony, in reality the latter often performs very well, and it generally allows a broader exploration of alternative trees than does maximum likelihood. Rapid methods of tree assembly, such as neighbour joining, come to the fore in the analysis of very large data sets of many sequences where other methods might prove too sluggish. Newly emerging methods such as spectral analysis and split decomposition emphasise visualising and exploring the data rather than simply finding the 'best' tree. Trees are hypotheses of evolutionary relationship, and are inferences made on limited data using often greatly simplified models of evolution that nevertheless pose immensely challenging computational problems. Finding the 'one true tree' is important, but so is knowledge of the range of signals in the data and the appropriateness of the model used.

6.10 Further reading

The sources given here are by no means exhaustive, but instead offer entry points into a large (and growing) literature on inferring evolutionary trees. Useful reviews include Swofford *et al.* (1996) and Penny *et al.* (1992). For a treatment of phenetics see Ridley (1986); Hull (1988) provides a stimulating overview of the history of the debates about phenetics and cladistics. The classic critique of distance methods is Farris (1981); Penny (1982) provides additional objections. Sober (1988) provides an excellent discussion of the foundations of parsimony and its relationship to maximum likelihood. Felsenstein (1981) has long been a proponent of maximum likelihood methods. For detailed evaluation of the merits of different likelihood models see Goldman (1993) and Yang (1994). For a technical treatment of the question of whether the maximum likelihood value for a given tree is unique see Steel (1994). Huelsenbeck and Rannala (1997) give a recent review of likelihood ratio tests. A good introduction to spectral analysis is given by Lento *et al.* (1995). Split decomposition is described by Bandelt and Dress (1992) and Dopazo *et al.* (1993). Hillis *et al.* (1992) describes their elegant work on experimental phylogenies using bacteriophage T7. Leitner *et al.* (1996) give an example of using a known phylogeny for HIV to evaluate different phylogenetic

methods. For a critique of experimental work see Sober (1993). Hillis's group have also made extensive simulation studies (Huelsenbeck and Hillis, 1993) and have recently applied parametric bootstrapping to phylogenetic studies (Huelsenbeck *et al.*, 1996). Sanderson (1995) reviews the debate about the merits of bootstrapping in phylogenetics.