Sequence data of DNA are becoming increasingly important within systematic and evolutionary biology. With this expanding emphasis, many investigators have rising expectations that sequence variation offers an almost unlimited potential to resolve phylogenetic relationships and thereby provide us with a highly robust, if not definitive, description of the pattern of life's history (e.g., Goodman et al., 1987; Sibley and Ahlquist, 1987; Goodman, 1989). Various reasons have been given in support of this expectation. One of the most common is that because DNA is the physical information system recording all inherited attributes of organismal history, direct comparison of its structure provides the most basic of all data for phylogenetic reconstruction. Inasmuch as all historical change is encoded in DNA, it is reasoned, it makes sense to examine that change directly. A second argument is one of large numbers: we can obtain far more character-state data from DNA sequences than we can, say, from traditional morphological comparisons, and therefore, statistically, as additional data are examined, true patterns of phylogenetic relationships will eventually emerge. Furthermore, it is often claimed that because morphology is so readily subject to convergence when compared to sequence data, obtaining more and more of the latter represents our best hope for resolving the phylogenetic pattern.

Technological innovations are also fueling the rush to obtain sequence data. Comparative sequence data are much easier and less expensive to collect than they were even a year ago. Just as importantly, there has been a rapid growth in the amount of computer hardware and software that makes the analysis of comparative sequence data readily accessible to workers at all levels of expertise (Platnick, 1989; see also many of the papers
in Doolittle, 1990). In particular, the availability of tree-building algorithms for microcomputers has been a contributing factor to the marked increase in the numbers of papers that use DNA sequences to answer phylogenetic questions (see summary in Swofford and Olsen, 1990). At its simplest, therefore, it is relatively straightforward to collect comparative sequence data, analyze them with a given tree-building method, and publish the results.

Yet, it is clear from the literature that phylogenetic inference using sequence data is anything but simple. There is, for example, substantial disagreement over the best method to resolve relationships from sequence data. And even if agreement could be reached on a general method, it still is not clear how that method might be applied in order to obtain a phylogenetic hypothesis that is judged reliable, or best, by some criterion. These and other problems make phylogenetic analysis of DNA sequences extremely complex.

There are three main classes of methods of phylogenetic inference applied to sequence data. First, with distance analysis, a distance (similarity/dissimilarity) coefficient is computed for all pairwise comparisons of raw sequence data, or for data corrected for multiple substitutions, and the taxa are then clustered based on the values in the distance matrix (Fitch and Margoliash, 1967; Farris, 1972; Felsenstein, 1984, 1988; Saitou and Nei, 1987). In general, the tree having the shortest overall branch length relative to that implied by the original data is chosen as the preferred tree. See Swofford and Olsen (1990) for details about optimization procedures, which differ among distance methods. Second, the method of maximum likelihood uses the original sequence data and, working from a prior evolutionary model of nucleotide substitution, computes a likelihood score for each tree given the original data (Felsenstein, 1981; Bishop and Friday, 1985; Saitou, 1988, 1990). That tree with the highest likelihood is the most preferred. The third method, cladistics or parsimony analysis, also uses the original sequence data, and from calculations of character-state changes (nucleotide substitutions) on alternative trees, the investigator chooses the one having the fewest character-state transformations, that is, the one for which homoplasies—parallelisms and reversals—are minimized (Fitch, 1971, 1977; Hendy and Penny, 1982; Cedergren et al., 1988).

There are several important reasons for preferring a parsimony approach to phylogenetic inference when using DNA sequences. First, parsimony methods are applied to the original sequence data rather than to some transformed distance metric, thus avoiding an inevitable loss of information (Farris, 1981, 1985; Penny, 1982). Second, while not entirely assumption free, parsimony methods are either less dependent upon assumptions about sequence evolution than are other methods; or they at least permit a more informed investigation of any assumptions that might be made (Cedergren et al., 1988). Third, algorithms that implement parsimony procedures are far more sophisticated than distance or maximum-likelihood alternatives and consequently allow a deeper analysis of the phylogenetic structure of
one's data and of the dynamics of sequence evolution. All three of these reasons mean that parsimony provides the investigator with enhanced interpretability of the results as compared to other methods.

Nevertheless, as noted above, the application of parsimony procedures to sequence data is complex and has not yet received sufficient attention. This chapter addresses two of the more general problems involving the use of parsimony procedures in phylogenetic inference: (1) given that there are physico-chemical/functional constraints on sequence evolution, especially in sequences coding for proteins or structural RNAs, how might parsimony be applied in order to infer phylogenetic relationships, and (2) how might we judge the phylogenetic informativeness of sequence data? Our inquiry into these problems will be based on an analysis of mitochondrial DNA (mtDNA), but the findings should be applicable to other parts of the genome as well.

Although it is a simple enough procedure to obtain a global parsimony estimate of the phylogenetic structure inherent in any given data set, recognition that rates of transitions are, at least in vertebrate mtDNA, much higher than those for transversions (Brown et al., 1982; Brown, 1983, 1985; Moritz et al., 1987), that there may be biases in base composition, and that silent substitutions in the third codon position are more common than those at first and second positions, have suggested to some investigators that parsimony procedures will have difficulties when used to resolve phylogenetic relationships, especially among more distantly related taxa (Moritz et al., 1987; Hayasaka et al., 1988a; Irwin et al., 1991). If homoplasy is too extensive, it is argued, obtaining a strongly corroborated phylogenetic tree might be difficult because even a minimum-length tree will have a high number of homoplastic changes (see Archie, 1989; Smith, 1989). To the extent that this is true, it raises several important questions: How might we undertake parsimony analysis so as to reduce, or at least identify, the influence of homoplasy? Second, how do we identify those parts of the tree that seem to be less strongly supported and at the same time specify which components of the data are less informative phylogenetically? Finally, in any given study, how do we determine how much sequence must be obtained in order to reveal a reliable phylogenetic signal?

These questions speak to the notion of reliability and informativeness, terms that have been used in different ways in the literature. Some investigators speak of "reliability of methods," referring to that method which finds the best-fit tree for a given set of data. Others view reliability in terms of an "ideal" result: the ability of a method, or a set of data, to enable us to infer a tree that reflects the one true phylogeny. It is well to keep the distinction of these two views of reliability in mind, for often it is implied that once a best-fit tree is obtained, so too has an accurate estimate of the one true phylogeny. It is sometimes argued, moreover, that under a particular model of nucleotide sequence change, one or more methods will fail to be reliable; that is, it would be expected that the best-fit tree generated by those methods will not be an accurate estimate of the
true phylogeny (e.g., Felsenstein, 1978). All of these views on reliability, however, are more arguments over methods, or perhaps of models of evolutionary change, inasmuch as the one true phylogeny can never be specified with certainty. Nor is it clear that a “realistic” evolutionary model can be developed without some prior hypothesis of relationships with which to investigate character-state change. Therefore, given only the data set of interest, it is difficult to say whether the best-fit tree produced by any method is or is not an accurate representation of the true phylogeny; it is simply the most economical explanation for the available data.

Ideally, “reliability” is best thought of as a parsimony problem and consequently should be judged in terms of congruence of data. “Informativeness,” likewise, is also a parsimony problem. The two concepts are related. A given tree, calculated to be the most-parsimonious, or best-fit, for some set of data, can be taken to be a reliable estimate of the true phylogenetic history of the taxa if most-parsimonious trees for other, independent data are congruent with it. A given sample of sequence can be judged phylogenetically informative if it corroborates the most-parsimonious tree derived from other data. This reciprocal duality between the structure of data and what is inferred from them manifests the application of parsimony (and congruence analysis) to hypotheses at two different levels of inference: that of the data (character homology or hypotheses of synapomorphy), and that of the phylogenetic hypotheses themselves (Cracraft and Mindell, 1989).

THE DATA

This study uses a single data set, and therefore, informativeness and reliability will be assessed in relation to the “stability” seen in the cladistic signal across partitioned subsets of the data. That is, to the extent to which clades are corroborated by these different subsets, the data will be judged informative and the phylogenetic hypotheses themselves reliable. “Noise,” on the other hand, will be judged to exist when the data do not corroborate a consistent (or stable) cladistic signal.

The sequence data analyzed in this study include an 898 base pair (bp) fragment of mtDNA that encompasses 458 bp of the 3’ end of the NADH-dehydrogenase subunit 4 (ND4 gene), 198 bp of three transfer RNA genes (tRNA^Hin, tRNA^Ser, and tRNA^Leu), and 242 bp of the 5’ end of the ND5 gene (Brown et al., 1982; Hayasaka et al., 1988a: fig. 1:629). Comparative sequences were available for 12 primate taxa: human (Homo sapiens), common chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), orangutan (Pongo pygmaeus), gibbon (Hylobates lar), Japanese macaque (Macaca fuscata), rhesus macaque (M. mulatta), crab-eating macaque (M. fascicularis), Barbary macaque (M. sylvanus), squirrel monkey (Saimiri sciureus), Philippine tarsier (Tarsius syrichta), and ring-tailed lemur (Lemur catta).
METHODS OF ANALYSIS

All analyses were undertaken using the Phylogenetic Analysis Using Parsimony (PAUP) computer program of D. L. Swofford (1990). Version 3.0 of this program for the Macintosh system is especially well suited for analysis of sequence data because it readily permits investigation of the effects of transitions versus transversions, the inclusion/exclusion of first, second, or third codon positions in coding sequences, and of nucleotide substitution parsimony of amino acid replacements. In addition, the program can compute consensus trees by different methods, perform bootstrap analyses, and calculate numerous measures of the information content for a given tree structure.

In order to investigate the effects of sample size of characters on the cladistic structure of the data, the 898 bp fragment was partitioned into four sequential subsets: (1) the first 225 bp of the ND4 gene; (2) a 458 bp segment including all of the ND4 gene that was sequenced; (3) a 656 bp fragment including the 458 bp ND4 fragment plus the 198 bp of the three tRNAs; and (4) the entire 898 bp fragment representing the ND4 fragment, tRNAs, and the 242 bp fragment of the ND5 gene. The effects of data set size were studied using this approach (rather than by random subsamples of different sizes) because systematists add sequence data sequentially and because it permitted an analysis of regions of the data set having different functional domains. In all analyses, the trees were rooted by designating the lemur as the outgroup. The results would have been the same if the tarsier had been designated the outgroup; therefore, this study is only investigating patterns of relationships within the clade consisting of the other primate species (the anthropoids; terminology of clades follows Hayasaka et al., 1988a). A nonprimate outgroup was not used in order to make our results more comparable to those of Hayasaka et al. (1988a).

Hayasaka et al. (1988a) proposed a phylogenetic hypothesis for the 12 primate taxa based upon substitution differences among those taxa and using the neighbor-joining method of Saitou and Nei (1987). A single tree was obtained (Fig. 10-1), which according to Hayasaka et al. (1988a:633), was also obtained using the unweighted pair-group method of distance analysis and by the distance-Wagner method. Hayasaka et al. concluded that “Because these three different methods give phylogenetic trees with the same topology, the phylogenetic relationships derived from these mtDNA sequence comparisons [see our Fig. 10-1] appear reliable.” It is not the purpose of this chapter to speak to the issue of primate relationships per se; rather the goal is to examine influences on the phylogenetic structure of these mtDNA sequences, to explore the implications of that structure for assessing the “reliability” of sequence data in general, and to investigate how parsimony analysis might be effectively applied to sequence data.

Two series of analyses are used to examine and describe the extent to which the primate mtDNA data set is phylogenetically informative. The first series employs parsimony analysis on different subsets of the data with the purpose of revealing the cladistic structure of most-parsimonious and
PARSIMONY AND PHYLOGENETIC INFERENCE

Figure 10-1 Phylogenetic hypothesis for 12 species of primates using a distance analysis (neighbor-joining method; Saitou and Nei, 1987) of an 898 bp fragment of mtDNA (from Hayasaka et al., 1988a). Names applied to clades discussed in this chapter follow the terminology of Hayasaka et al. (1988a): (1) hominines, (2) great apes, (3) hominoids, (4) catarhines, (5) anthropoids, (6) macaques, and (7) prosimians.

near-most-parsimonious trees. The most-parsimonious tree, or collection of equally parsimonious trees, was found. Once its length was known, all trees within five steps of that length were retained, and a majority-rule consensus tree (Margush and McMorris, 1981) computed (majority-rule consensus trees were adopted because they permit an analysis of the relationship between the amount of phylogenetic signal and data set size and because they are comparable to the results of bootstrap analyses as discussed below). The collection of trees having lengths within 1% of the minimum-length tree was also examined and will be mentioned briefly.

A second series of analyses used the same subsets of data and constructed a majority-rule bootstrap tree (Felsenstein, 1985) from 100 randomly sampled (with replacement) replications of each of the subsets. The bootstrap, in this case, is not intended as a test of statistical significance but as an additional means of describing and evaluating the structure of the phylogenetic signal inherent in the data.

RESULTS

Global Parsimony Analysis

Most-Parsimonious and Near-Most-Parsimonious (Five-Step) Trees

A global parsimony analysis was undertaken using all nucleotide substitutions in each of the four subsets of the 898 bp fragment. This analysis was designed to answer two principle questions: Can a global parsimony analysis, in which transitions and transversions are equally (uniformly)
weighted, reveal a stable phylogenetic signal? And, second, does the sta-
bility of that signal change as sample size is increased? Most-parsimonious
trees are described but not figured. Primary attention is given to the set
of trees within five steps of the most-parsimonious tree (or trees), but the
collection of trees within 1% of the minimum-length tree is also examined.

1. ND4:225 bp Fragment. A parsimony analysis of the smallest data set
of 225 characters yielded six equally parsimonious trees of 264 steps. These
trees have only two topological ambiguities: (1) the hominine trichotomy;
and (2) a trichotomy among the hominoids, macaques, and the squirrel
monkey. All other relationships were as postulated by Hayasaka et al.
(1988: fig. 3:636) (see Fig. 10-1).

There are 95 trees within five steps of the shortest trees. The consensus
tree exhibits relatively poor structure except within the macaques (Fig.
10-2A). There is no resolution within the great apes clade or among the
basal lineages of the anthropoids.

2. ND4:458 bp Fragment. When the analysis is expanded to include all of
the ND4 fragment, a single most parsimonious tree of 608 steps is found
that is similar to the distance tree except for a sister-group relationship
between the chimp and gorilla.

In the 458 bp data set for ND4 there are 17 trees within five steps of
the single most-parsimonious tree (Fig. 10-2B). Resolution within the homi-
noinds has improved substantially, although the human-chimpanzee-gorilla
trichotomy, while recognized in 88% of the trees, is not resolved internally.
The hominoid relationships of the gibbon are still somewhat ambiguous,
occurring in only 53% of the trees. The additional data do not change the
resolution within the macaques, but do now place them closer to the homi-
noinds than to the squirrel monkey in 65% of the trees.

3. ND4 + tRNAs: 656 bp Fragment. A parsimony analysis of the ND4
fragment and the three tRNA genes produced a single most-parsimonious
tree of 766 steps identical to that of the preceding analysis in which the
gorilla is joined to the chimpanzee.

There were 9 trees within five steps of the most-parsimonious tree (Fig.
10-2C). With the addition of the 198 bp tRNA fragment to the ND4 se-
quence, resolution within the hominoid clade is improved, although the
human-chimpanzee-gorilla trichotomy remains. Resolution in other parts
of the tree remains relatively the same as for the 458 bp fragment.

4. ND4 + tRNAs + ND5: 898 bp Fragment. A parsimony analysis of the
entire data set yielded two equally most-parsimonious trees (1,157 steps),
one uniting chimpanzee and human, the other uniting chimpanzee and
gorilla; all other relationships were as in the distance tree (Fig. 10-1).

Within five steps of the most-parsimonious trees, five trees were found.
The consensus tree (Fig. 10-2D) is fully resolved but a human-chimpanzee relationship was supported in only three of those trees (60%). Interestingly, only the branch uniting the orangutan to the hominines received less support as compared to the 656 bp fragment five-step consensus tree. Overall there was a marked improvement in resolution with the larger data set.

5. ND4 + ND5 Fragments Compared With the tRNA Fragment. In order to assess the relative influence on phylogenetic structure of mtDNA sequences that code for proteins versus structural RNAs, the two regions were analyzed separately. A parsimony analysis of the 700 bp encompassing the two coding regions of ND4 and ND5 taken together produced two equally parsimonious trees (999 steps) that differ only in the relationships of the chimpanzee to humans or to the gorilla; in all other respects the two trees are identical to the distance tree (Fig. 10-1).

There were five trees within five steps of the two shortest trees, and their consensus is shown in Figure 10-2E. Resolution is improved over the 656 bp fragment except in the placement of the orangutan with the hominines. Compared to the entire 898 bp data set (Fig. 10-2D), the five-step tree of ND4 + ND5 has less resolution within the hominines but the relationship between the hominoids and the macaques is more strongly supported by the ND4 + ND5 data set.

An analysis of the 198 bp fragment that includes the three tRNAs suggests that their contribution to instability is minimal. A single most-parsimonious tree was found (154 steps) that resolved relationships within the hominoids and macaques identical to those of the distance tree. This tRNA tree, however, placed the macaques outside the other taxa to yield: (macaques (tarsier (squirrel monkey + hominoids))).

There are 254 trees five steps away from the shortest tRNA tree and, not unexpectedly, the consensus tree (Fig. 10-2F) exhibits relatively little resolution.

Near-Most (1%)-Parsimonious Trees

The collection of trees within 1% of the length of the most-parsimonious tree has been used as a method to investigate the cladistic structure of sequence data (Smith, 1989) and was also examined in this study. Table 10-1 summarizes the results of this analysis.

The number of trees recovered within 1% of the shortest tree(s) did not vary noticeably among data sets despite the fact that larger data sets had trees of longer branch lengths. Only two clades, hominines and hominoids, consistently gained cladistic support as the size of the data set increased. All others showed decreasing support or no marked change. Some clades (anthropoids and macaques) were well defined no matter what size the data set. Interestingly, a global parsimony analysis shows strong resolution within the macaques across the four primary data sets, but no resolution is seen within the hominines.
A. ND4: 225 bp
95 trees (269 steps or less)
shortest tree = 264 steps (5 trees)

B. ND4: 458 bp
17 trees (613 steps or less)
shortest tree = 606 steps (1 tree)

C. ND4+tRNAs: 656 bp
9 trees (771 steps or less)
shortest tree = 766 steps (1 tree)

D. ND4+tRNAs+ND5: 898 bp
5 trees (1162 steps or less)
shortest tree = 1157 steps (2 trees)
Figure 10-2  Majority-rule consensus trees of a global-parsimony analysis of all trees within five steps of the most-parsimonious solution for subsets of the ND4 + tRNAs + ND5 fragment of mtDNA, as discussed in the text. Numbers next to internal branches represent frequencies of occurrence for those clades among the most- and near-most-parsimonious solutions.
<table>
<thead>
<tr>
<th>Fragment (no. of trees)</th>
<th>Human + Chimpanzee</th>
<th>Hominines</th>
<th>Great Apes</th>
<th>Hominoids</th>
<th>Catarrhines</th>
<th>Anthropoids</th>
<th>Japanese + Rhesus</th>
<th>Japanese + Rhesus + Crab-eating</th>
<th>Macaques</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND4: 225 bp (29)</td>
<td>n.r. *</td>
<td>52</td>
<td>100</td>
<td>55</td>
<td>62</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>ND4: 458 bp (24)</td>
<td>n.r.</td>
<td>79</td>
<td>96</td>
<td>58</td>
<td>71</td>
<td>100</td>
<td>92</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>ND4 + tRNAs: 656 bp (22)</td>
<td>n.r.</td>
<td>86</td>
<td>95</td>
<td>73</td>
<td>64</td>
<td>100</td>
<td>100</td>
<td>82</td>
<td>100</td>
</tr>
<tr>
<td>898 bp fragment (17)</td>
<td>n.r.</td>
<td>100</td>
<td>59</td>
<td>100</td>
<td>59</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td>100</td>
</tr>
<tr>
<td>ND4 + ND5: 700 bp (13)</td>
<td>n.r.</td>
<td>100</td>
<td>69</td>
<td>100</td>
<td>69</td>
<td>100</td>
<td>85</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>tRNAs: 198 bp (25)</td>
<td>56</td>
<td>100</td>
<td>n.r.</td>
<td>100</td>
<td>n.r. †</td>
<td>60</td>
<td>100</td>
<td>52</td>
<td>100</td>
</tr>
</tbody>
</table>

*n.r., clade not resolved at 50% level.

†Percentage of trees in which clade is resolved in majority-rule consensus trees.

‡Squirrel monkey clustered with the hominoids.
**Summary of Global Parsimony Analysis**

Numerous cladistic groupings are consistently revealed in a global parsimony analysis. With the exception of the smallest data set, the first 225 bp fragment of ND4, global parsimony analysis of each of the other data sets was able to resolve all relationships consistently except for the trichotomy within the hominines. This is also true of the combined ND4 + ND5 fragment of 700 bp, and, as noted above, even the most-parsimonious tree of the tRNAs was able to resolve most of the relationships.

These results indicate that this relatively small 898 bp fragment of mtDNA contains significant phylogenetic signal for taxa whose divergences extend as far back as 50 to 60 million years, that even "noise" produced by the high transition rate of mtDNA cannot obscure the signal, and that a global parsimony analysis is sufficiently powerful to resolve a consistent cladistic pattern even in data that are generally thought to be "noisy."

These conclusions are reinforced when the collections of near-most-parsimonious trees are examined (Fig. 10-2). In general, there is improved resolution as more data are added (Fig. 10-2A through 10-2E), but the trend is not always strong. Most importantly, a consistently strong cladistic signal is found in all but the very smallest data sets.

When the structure of trees within 1% of the most-parsimonious solution is examined, however, far fewer of the relationships are identified by a strong cladistic signal (e.g., clusters supported in 90% or more of the trees; see Table 10-1). Only the monophyly of the macaques and of the anthropoids is consistently revealed across all four primary data sets. As noted earlier, furthermore, saving 1% of the most-parsimonious trees does not produce a clear convergence in the phylogenetic signal as the data set is enlarged. This results from the fact that in each subset of the data there are many trees within 1% of the minimum-length tree, including some of extremely variant tree topology.

**Transversion Parsimony Analysis**

*Most-Parsimonious and Near-Most-Parsimonious (Five-Step) Trees*

Does the large bias for transition substitutions within mtDNA increase the "noise" and lead to instability of the phylogenetic signal within the primate data set? Can the signal be improved by examining transversions alone? And, does the stability of a phylogenetic signal inferred from a parsimony analysis of transversions improve with increasing sample size? These questions were investigated by repeating the above analysis using transversion differences.

1. **ND4: 225 bp Fragment.** An analysis of the first 225 bp of the ND4 fragment produced two equally parsimonious trees (79 steps). In both trees, the macaques were not fully resolved, with only the Japanese-rhesus macaque clade being held in common. The major difference between the two
trees was found in the placement of *Homo*: in one tree it was with gorilla, in the other with chimpanzee.

There were 681 trees within five steps of the most-parsimonious solutions. Their majority-rule consensus (Fig. 10-3A) shows very strong support for three clades, the anthropoids, catarrhines, and macaques, but fails to resolve the relationships of any other group.

2. *ND4*: 458 bp Fragment. A transversion parsimony analysis of the entire ND4 fragment yielded two equally parsimonious trees (196 steps). One is identical to the distance tree (Fig. 10-1), and the other failed to resolve a trichotomous relationship at the base of the macaques: (Barbary, crab-eating, Japanese + rhesus).

Although 114 trees were found within five steps of the most-parsimonious solution, the majority-rule consensus (Fig. 10-3B) shows a marked improvement in resolution compared to the 225 bp fragment. Clades are resolved within the hominoids, but not within macaques.

3. *ND4* + tRNAs: 656 bp Fragment. A transversion analysis of the ND4 fragment along with the tRNAs produced two equally parsimonious trees (239 steps) that have identical structure to the two most-parsimonious trees of the preceding analysis.

There were 59 trees within five steps of the two most-parsimonious trees. Compared to the 458 bp fragment, some additional, slight resolution is apparent in the hominines and the macaques (Fig. 10-3C).

4. *ND4* + tRNAs + *ND5*: 898 bp Fragment. A transversion parsimony analysis of the entire data set yielded a single most-parsimonious tree (381 steps). It has the same structure as the distance tree (Fig. 10-1).

For the transversions in the total data set, 46 trees were found within five steps of the most-parsimonious tree. Although there is no resolution within the hominines or macaques, compared to the previous consensus trees, there is a marked increase in support for the hominines and catarrhines (Fig. 10-3D).

5. *ND4* + *ND5* Fragments Compared With the tRNA Fragment. In a transversion parsimony analysis of the ND4 and ND5 fragments combined (700 bp), a single most-parsimonious tree was obtained (338 steps) that is identical to the distance tree. A transversion parsimony analysis of the three tRNAs (198 bp) found seven equally parsimonious trees (42 steps). The consensus tree for these showed the chimpanzee united with the human in 77% of the cases, but with no resolution among the other hominoids, including the gorilla. The Japanese and rhesus macaques were clustered as sister-species, but no other nodes were resolved within macaques, and the squirrel monkey was united with the macaques.

Transversion analysis of the 700 bp coding region finds 66 trees within five steps of the most-parsimonious tree. Their majority-rule consensus
tree (Fig. 10-3E) is very similar to that of the entire 898 bp data set (Fig. 10-3D) except that the great apes are much less supported (76% versus 91%). Not unexpectedly, there are a large number of trees within five steps of the most-parsimonious solutions for the tRNA fragment. The majority-rule consensus for 3,000 trees (computer memory was exhausted; many more could have been found) is shown in Figure 10-3F. Only the macaques as a group are well defined.

Near-Most (1%)-Parsimonious Trees

Table 10-2 summarizes majority-rule consensus tree support for all trees with lengths within 1% of the most-parsimonious transversion solutions. A consistent pattern is observed: except for the tRNA fragment, virtually all of the "deep" cladistic events are strongly supported in each of the five primary analyses, and this also includes the shortest 225 bp fragment. In contrast, strong resolution of the "more recent" cladistic events, including those within the macaques and hominines, is lacking.

Summary of Transversion Parsimony Analysis

The most general conclusion to be drawn from a comparison of the global parsimony analysis (Fig. 10-2; Table 10-1) with the transversion parsimony analysis (Fig. 10-3; Table 10-2) is that the use of transversions, by themselves, reveals a strong cladistic signal for the relatively older clades. This is especially evident when examining those trees having a length within 1% of the most-parsimonious solutions (Table 10-2), which in most of the data sets included all those trees within two or three steps of the minimum-length solution. Since many more trees were examined in the five-step analysis (Fig. 10-3), support for some of these deeper branches declined, but in general remained fairly strong. None of these methods of analysis, when applied to transversion differences alone, were able to reveal a strong cladistic signal within the hominines or within the macaques.

The five step analysis revealed a slight increase in the strength of the cladistic signal as the data set increased in size. The 656 bp fragment (Fig. 10-3C) is not much more informative than the 458 bp fragment (Fig. 10-3B), but the total data set (Fig. 10-3D) and the ND4 + ND5 fragment (700 bp; Fig. 10-3E) exhibit stronger cladistic support within the hominoids.

Global Parsimony of First and Second Codon Positions

Does the high rate of transitions in third positions contribute noise and therefore result in decreased stability of the phylogenetic signal? This question was investigated by a global parsimony analysis of the ND4 and ND5 fragments in which the third positions were excluded. The analysis yielded a single most-parsimonious tree of 519 steps. There were only two differences from the distance tree. The first had the chimpanzee united with the gorilla, but the most unexpected result (given the preceding anal-
A. ND4: 225 bp
   661 trees (84 steps or less)
   shortest tree = 79 steps (2 trees)

B. ND4: 458 bp
   114 trees (201 steps or less)
   shortest tree = 198 steps (2 trees)

C. ND4+tRNAs: 656 bp
   59 trees (244 steps or less)
   shortest tree = 230 steps (2 trees)

D. ND4+tRNAs+ND5: 898 bp
   46 trees (388 steps or less)
   shortest tree = 381 steps (1 tree)
E. ND4+ND5: 700 bp
66 trees (343 steps or less)
shortest tree = 338 steps (1 tree)

F. tRNAs: 198 bp
3000+ trees (47 steps or less)
shortest tree = 42 steps (7 trees)

Figure 10-3  Majority-rule consensus trees of a transversion parsimony analysis of all trees within five steps of the most-parsimonious solution for subsets of the ND4 + tRNAs + ND5 fragment of mtDNA, as discussed in the text. Frequencies as in Figure 10-2.
<table>
<thead>
<tr>
<th>Fragment (no. of trees)</th>
<th>Human + Chimpanzee</th>
<th>Hominines</th>
<th>Great Apes</th>
<th>Hominoids</th>
<th>Catarrhines</th>
<th>Anthropoids</th>
<th>Japanese + Rhesus</th>
<th>Japanese + Rhesus + Crab-eating</th>
<th>Macaques</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND4: 225 bp (17)</td>
<td>n.r.*</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>53</td>
<td>n.r.</td>
<td>100</td>
</tr>
<tr>
<td>ND4:458 bp (14)</td>
<td>n.r.</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>n.r.</td>
<td>n.r.</td>
<td>100</td>
</tr>
<tr>
<td>ND4 + tRNAs: 656 bp (8)</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>n.r.</td>
<td>n.r.</td>
<td>100</td>
</tr>
<tr>
<td>898 bp fragment (27)</td>
<td>59</td>
<td>100</td>
<td>96</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>n.r.</td>
<td>n.r.</td>
<td>100</td>
</tr>
<tr>
<td>ND4 + ND5: 700 bp (38)</td>
<td>n.r.</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>n.r.</td>
<td>n.r.</td>
<td>100</td>
</tr>
<tr>
<td>tRNAs: 198 bp (77)</td>
<td>77</td>
<td>n.r.</td>
<td>n.r.</td>
<td>74</td>
<td>n.r.*</td>
<td>100</td>
<td>64</td>
<td>n.r.</td>
<td>100</td>
</tr>
</tbody>
</table>

*n.r., clade not resolved at 50% level.

1Percentage of trees in which clade is resolved in majority-rule consensus trees.

2Squirrel monkey clustered with macaques.
yses) was the closer relationship of the gibbon, rather than the orangutan, to the great apes.

There were eight trees within five steps of the shortest tree (Fig. 10-4A). The consensus tree showed little resolution within either the hominoids or the macaques, but it did resolve the monophyly of both of these clades and clustered them relative to the squirrel monkey. This analysis suggests that attempting to enhance the phylogenetic signal by eliminating transitions in the third position was unsuccessful. It had no effect on improving the resolution of the older cladistic events, which were strongly resolved by global parsimony, and not unexpectedly, it led to a decrease in the signal within the macaques (compare Fig. 10-4A with Fig. 10-2E).

Amino Acid Replacements: Substitution Parsimony

Amino acid replacements are often taken as being more informative of relationships than nucleotide substitutions, especially for more distant branching events. The structure of the phylogenetic signal of replacements was examined by parsimony analysis of the numbers of substitutions that account for amino acid differences among the taxa. A transformation matrix for vertebrate mtDNA provided by D. L. Swofford was used to convert amino acid differences to substitution differences and then each tree was evaluated.

A single most-parsimonious tree of 387 steps was found for the 231 codons of the ND4 and ND5 fragments arranged tandemly. Relationships were similar to those of the distance tree except within the hominoids where gorilla and human were united and the gibbon (not orangutan) was the sister-group of the hominines.

There were 49 trees within five steps of the most-parsimonious tree. The majority-rule consensus tree (Fig. 10-4B) shows relatively little resolution, either within the macaques or the hominoids, thus suggesting that this method is less effective in revealing phylogenetic signal than is a global parsimony analysis.

Bootstrap Analysis

Global Parsimony Bootstrap

1. **ND4:** 225 bp Fragment. A bootstrap analysis of the first 225 bp of ND4 produced a majority-rule consensus tree identical to the distance tree except for chimpanzee being united with gorilla (Fig. 10-5A). The latter relationship means little, however, given that all three alternatives within the hominines are about equally supported. Clades having substantial support in this data set include: (1) the macaques and their inclusive groupings; (2) the great apes; and (3) the anthropoids.

2. **ND4:** 458 bp Fragment. A bootstrap analysis of the entire ND4 fragment produced a tree very similar to the preceding, except that human
Figure 10-4  (A) A majority-rule consensus tree of a global parsimony analysis of all trees within five steps of the most-parsimonious solution for the tandemly arranged fragments of ND4 and ND5 genes of mtDNA with third codon positions excluded. (B) A majority-rule consensus tree of a global parsimony analysis of nucleotide substitutions resulting in amino acid replacements in 231 codons of the ND4 and ND5 genes of mtDNA. Frequencies in both (A) and (B) as described in Figure 10-2.
and chimpanzee are narrowly united (Fig. 10-5B). Support for each of the clades has generally improved, especially for the catarrhines, the hominines, and the macaques.

3. ND4 + tRNAs: 656 bp Fragment. The majority-rule consensus tree for this data set (Fig. 10-5C) is identical to the preceding. In general, support for each clade increases, although the catarrhine clade is found in only 50% of the replicates.

4. ND4 + tRNAs + ND5: 898 bp Fragment. The majority-rule consensus tree for all the data keeps the same pattern of relationships as the preceding, except for uniting chimpanzee and gorilla (50% of the replicates; Fig. 10-5D). Bootstrap support for the other clades has increased, sometimes substantially (e.g., hominoids, catarrhines).

5. ND4 + ND5 Fragments Compared With the tRNA Fragment. A bootstrap majority-rule consensus tree of the protein-coding region (Fig. 10-5E) shows a similar pattern to that of the distance tree. Levels of support are close to those of the 898 bp tree (Fig. 10-5D). Bootstrapping the 198 bp of the rRNAs (Fig. 10-5F) suggests that they contribute less noise to the data than might be expected, except at the base of the anthropoids where the squirrel monkey is united with the hominoids. The tRNAs, however, strongly support the monophyly of the macaques and a rhesus + Japanese macaque relationship.

Summary of the Global Parsimony Bootstrap

Not unexpectedly, the phylogenetic signal, as measured by the percentage of congruent bootstrap replicates, improves as the sample size increases. Thus, for the 898 bp fragment (Fig. 10-5D), six clades were supported more than 97% of the time. Nevertheless, if one preferred to interpret the "significance" of these results stringently by collapsing all branches that failed to meet a 90 or 95% criterion, the bootstraps of characters for each subset of the data would reveal a cladistic structure not very much different from that seen in previous analyses (Fig. 10-2).

Transversion Parsimony Bootstrap

1. ND4: 225 bp Fragment. A bootstrap of transversions contained in the first 225 bp of the ND4 fragment produces a majority-rule consensus tree much like the distance tree except for the lack of resolution at the base of the macaques (Fig. 10-6A). The monophyly of the macaques, the catarrhines, and the anthropoids is strongly supported by a bootstrap analysis of these data.

2. ND4: 458 bp Fragment. A bootstrap of the entire ND4 fragment produced a tree identical to the distance tree (Fig. 10-6B). Relative to the 225 bp fragment, the pattern of support varies. Clades within the hominoids are better supported with this expanded data set, although the monophyly
Figure 10-5  Majority-rule bootstrap trees of a global parsimony analysis for subsets of the ND4 + tRNAs + ND5 fragment of mtDNA, as discussed in the text. Frequencies of occurrence among the bootstrap trees are presented for the internal branches.
Figure 10-6  Majority-rule bootstrap trees of a transversion parsimony analysis for subsets of the ND4 + tRNAs + ND5 fragment of mtDNA, as discussed in the text. Frequencies as in Figure 10-5.
of the catarrhines is much less so. Support for a human + chimpanzee relationship has increased to 66% of the replicates.

3. ND4 + tRNAs: 656 bp Fragment. The relationships implied by this data set (Fig. 10-6C) are the same as in the preceding analysis. There is an increase in the support of all clades except that of the catarrhines.

4. ND4 + tRNAs + ND5: 898 bp Fragment. A bootstrap of the transversions in the entire data set yielded a tree showing no changes in cladistic structure from the preceding analysis, but with a general increase in the cladistic support at most nodes (Fig. 10-6D).

5. ND4 + ND5 Fragments Compared to the tRNA Fragment. A transversion bootstrap analysis of the combined ND4 and ND5 sequences once again produced the same tree as did using all the data and, in general, with approximately the same level of support (Fig. 10-6E). An analysis of transversions in the tRNAs shows that this data set retains much of the phylogenetic structure present in the protein-coding regions, but virtually all of the nodes are poorly supported (Fig. 10-6F). In the tRNA data set, the squirrel monkey clusters with the macaques, showing ambiguity at the base of the anthropoids.

Summary of the Transversion Bootstrap Analysis
As with the global parsimony bootstrap analysis (Fig. 10-5), there was a general increase in support for the nodes as sample size was increased. Neither bootstrap analysis provided significantly better support than the other. Although the transversion bootstrap produced much better resolution within the hominines, the global bootstrap analysis resolved the relationships within the macaques much more convincingly. Both did about equally well with the other parts of the tree.

The trees of Figure 10-6 are all majority-rule consensus trees, and if the level of stringency of cladistic support was taken to be 90 or 95% of the bootstrap replicates, the resulting topologies would not differ very much from those of the transversion parsimony analysis (Fig. 10-3).

DISCUSSION

Informativeness, Noise, and the Primate Data Set
A major concern when analyzing any systematic data is to assess the extent to which they are phylogenetically informative. Systematists are also interested in knowing whether a particular analytical method is effective in maximizing informativeness and thus results in phylogenetic hypotheses that are more strongly supported than those produced by other methods. These problems are especially acute with sequence data, not only because
of the large numbers of characters that are normally generated, but also because the evolutionary dynamics of DNA are such that a large amount of noise (i.e., homoplasy) would not be unexpected.

How might the informativeness and noisiness of the primate data set be evaluated? One approach might examine sets of near-most-parsimonious trees and judge informativeness by the degree to which the cladistic structure is corroborated across those trees. A second method might examine consensus trees produced by resampling of the data themselves and ask to what extent each node is supported; that is, in what percentage of the trees forming the consensus does a given node occur? Both methods examine the structure of data, but in different ways.

A consideration of the amount of support for each node as seen in the majority-rule consensus trees (Figs. 10-2 and 10-3; Tables 10-1 and 10-2) permits a detailed description of the phylogenetic signal contained in the data. In the global majority-rule consensus of trees within five steps of the most-parsimonious solutions (Fig. 10-2), the relationships within the macaques are fully resolved using the largest data sets. The hominines always cluster relative to the orangutan, and these two groups cluster relative to the gibbon. The major lineages of the tree—hominoids, catarrhines, anthropoids, and macaques—are also well supported by the largest data sets. In the transversion majority-rule trees (Fig. 10-3), there is little or no phylogenetic signal within the macaques or hominines, but the data unambiguously resolve all other clades, especially using the largest data sets. Taken together, these two analyses reveal a consistent phylogenetic signal for all clades except those within the hominines.

Bootstrap analysis constitutes a second approach to characterizing the amount of phylogenetic signal in the data. Bootstrap sampling of the total data set (Fig. 10-5) reveals a strong phylogenetic signal within the macaques, and comparison with the bootstrap analysis of transversions alone (Fig. 10-6) leads to the conclusion that a substantial component of that signal is in fact coming from transition differences. Within the hominines, however, it is transversions, not transitions, that provide the most consistent signal. In both bootstrap analyses, the major lineages are all well supported using the largest data sets, although the phylogenetic position of the squirrel monkey is the least well-corroborated.

How do the two approaches compare? Not unexpectedly, the results using these two different methods parallel one another to a very great extent (e.g., compare Figs. 10-2 and 10-5; and 10-3 and 10-6). Whenever a clade is strongly supported in a near-most-parsimonious consensus tree, it is also strongly supported in the comparable bootstrap tree. Differences arise in those clades that lack strong support. Because bootstrap trees tend to be fully resolved, they have the appearance of being more informative, yet there is generally little character support underlying that increased resolution. Still, when examination of the structure of near-most-parsimonious trees often yields an unresolved polytomy, bootstrap resampling of the data seems more sensitive to finding the “correct” signal, even
if that signal is not especially strong. Again, one need not interpret bootstrap results in a strict statistical sense, but simply as another method of evaluating the presence and strength of signal in the data. Thus, bootstrapping of transversions across the entire data set provides resolution of relationships within the macaques (Fig. 10-6D), whereas the majority-rule consensus analysis of near-most-parsimonious trees (Fig. 10-3D) does not. This result generally holds for most subsets of the data, not only for relationships within the macaques, but also within the hominines.

Does Phylogenetic Signal Stabilize with More Data?

A critical practical problem of phylogenetic inference using DNA sequences is deciding how much sequence must be obtained before relationships can be resolved. In attempting to answer this question, factors other than sequence length per se cannot be ignored. It is generally assumed, for example, that certain segments of DNA will be more informative than others and therefore less sequence will be required to reveal a phylogenetic signal. This supposition usually entails considerations of evolutionary rate, which is discussed in the next section. Here we ask whether phylogenetic signal within the primate data set is related to data set size and which method of analysis best reveals that signal.

The global parsimony analysis of near-most-parsimonious trees shows an overall increase in the strength of signal as sample size is increased (Fig. 10-2A through 10-2E). There is a marked improvement in the 458 bp tree as compared to the 225 bp tree. With respect to the 656 bp consensus tree (Fig. 10-2C), some nodes are slightly better supported compared to the 458 bp tree, others are not. Generally, the 898 bp and 700 bp trees have stronger corroboration than the 656 bp tree.

The transversion parsimony analysis of near-most-parsimonious trees also exhibits increasing strength of signal with increasing sample size (Fig. 10-3). That increase in signal, however, is confined to the deeper branches and to the hominoids. Resolution within the macaques is not affected by increases in sample size using transversions alone.

As noted earlier there is a general increase in phylogenetic signal, as determined by bootstrap support, when sample size increases. This is true for both the global parsimony and transversion parsimony analyses (Figs. 10-5, 10-6). Not all nodes show increased support between successive subsets of data, and not all nodes are well supported, but there is a definable convergence toward a single tree. These results suggest that bootstrap analysis may be somewhat more informative as a descriptor of the effects of sample size on phylogenetic signal than is examination of near-most-parsimonious trees.

Evolutionary Rate and Phylogenetic Signal

The degree to which a segment of DNA is considered likely to be informative is generally thought to be related to its rate of base substitution
relative to the divergence times that the investigator is trying to resolve. Segments with high rates of change are assumed to be less informative for deeper branches because the accumulation of homoplasy (parallel and back mutations) erases any signal (Hayasaka et al., 1988a; Smith, 1989). As substitutions "saturate," genetic distances among the taxa approach an asymptote relative to divergence time (Brown, 1983; Moritz et al., 1987), and this observation has led to the conclusion that mtDNA sequence variation will be less useful in resolving relationships among relatively more distantly related taxa (Moritz et al., 1987).

To the extent that data are available (Brown, 1985), the genes examined in this study (ND4 and NOS) might be categorized as "relatively fast" compared to many other protein-coding genes, tRNAs, or ribosomal RNA genes. Hayasaka et al. (1988a:637–638) estimated that the relationship between divergence time and the number of substitutions among taxa [corrected for multiple mutations (Gojobori et al., 1982)] is linear only up to about 30 million years and that older divergence times (e.g., those deeper than the base of the hominoids) cannot be resolved "conclusively." Yet, such a conclusion is method-dependent. Even though ND4 and NOS are "fast," the present analysis suggests their rate may have only marginal influence on the relationship between phylogenetic signal and divergence time (at least within primates). Depending upon how parsimony is applied to the data, these genes are just as capable of resolving the more ancient divergences within the primates as they are the most recent. Thus, by examining the relative influence of transitions and transversions, a reasonably consistent phylogenetic signal is revealed across the tree (relationships within the hominines constitute a slight exception; see below). These conclusions deserve to be tempered, however, by noting that transition/transversion differences do not become equal until the very most distant relationships between the prosimians and the other taxa are considered (Hayasaka et al., 1988a; see our Table 10-3). It could be argued, therefore, that relationships within the anthropoids are not affected by a saturation effect (although saturation is evident at the base of the anthropoids).

**The Resolving Power of Most-Parsimonious Trees**

It is often thought within molecular systematics that parsimony methods will often fail to be informative because of the high degree of homoplasy in sequence data (see next section). Therefore, it is useful to examine that supposition by asking whether the most-parsimonious trees, based on a consideration of all of the data, were in fact accurate estimates of the phylogenetic relationships of the primates, or whether the "noise" in these "fast" evolving genes made that resolution difficult or impossible.

The answer to this question may be surprising to some: the relationships of the primates were consistently revealed by the most-parsimonious trees of each of the data sets. In general, the only ambiguity involved resolution
<table>
<thead>
<tr>
<th>Clade</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Human</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2. Chimpanzee</td>
<td>74/5</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3. Gorilla</td>
<td>84/8</td>
<td>86/9</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4. Orangutan</td>
<td>108/35</td>
<td>119/34</td>
<td>116/23</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5. Gibbon</td>
<td>116/45</td>
<td>124/44</td>
<td>123/45</td>
<td>115/52</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6. Japanese macaque</td>
<td>134/74</td>
<td>144/73</td>
<td>139/72</td>
<td>143/75</td>
<td>145/77</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7. Rhesus macaque</td>
<td>131/77</td>
<td>145/76</td>
<td>135/75</td>
<td>141/78</td>
<td>133/80</td>
<td>32/3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8. Crab-eating macaque</td>
<td>149/74</td>
<td>166/73</td>
<td>162/72</td>
<td>158/75</td>
<td>154/77</td>
<td>71/4</td>
<td>81/5</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9. Barbary macaque</td>
<td>155/74</td>
<td>149/73</td>
<td>146/72</td>
<td>137/77</td>
<td>139/77</td>
<td>103/8</td>
<td>100/9</td>
<td>104/6</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
of relationships within the hominines, which has been a long-standing prob-
lem that even very large amounts of data have had difficulty in resolving.
The most parsimonious solution of the smallest data set, the 225 bp frag-
ment, also showed a trichotomy at the base of the anthropoids, yet was
still remarkably effective in resolving relationships among the other taxa.

It therefore seems that a straightforward parsimony analysis of un-
weighted data, including both transitions and transversions, is an effective
method for estimating what appears to be the "correct" tree, even in subsets
of the data. It may be that because most of the data are not saturated with
transition substitutions, there remains relatively little noise in the data and,
therefore, parsimony is effective in revealing the signal. On the other hand,
it may be that a simple parsimony approach is in fact a very powerful
method and can discern phylogenetic signal, even in data having a strong
component of noise [consistency indexes (Kluge and Farris, 1969) calcu-
lated after excluding uninformative characters typically had values slightly
less than 0.600].

Parsimony and Phylogenetic Inference Using DNA Sequences

Many workers have long advocated parsimony analysis as a method of
phylogenetic inference for DNA sequence data (Fitch, 1971, 1977; Hendy
and Penny, 1982). At the same time, there has been concern within the
molecular systematic community that once the algorithmic problems have
been solved—and large advances have been made in recent years—there
will still remain important questions as to exactly how parsimony proce-
dures are to be applied to sequence data. The focus of most applications
of parsimony has been to find the most-parsimonious tree or trees. Yet,
systematic (comparative) and functional analyses of DNA and its evolution
have established a body of knowledge that creates a broad spectrum of
choices and challenges for the systematist. Included in this body of knowl-
edge are the observations that the frequency of transitions and transver-
sions are not always equal; third position changes are more frequent than
first or second; and secondary structure configurations can bias the fre-
quency and direction of nucleotide change.

The investigator is thus faced with procedural difficulties from the outset,
difficulties that raise important philosophical and empirical questions. At
their most basic level, these questions involve the problem of what con-
stitutes evidence in phylogenetic inference, a subject that has received too
little attention in the literature (a recent notable exception is Kluge, 1989).
Do we base our best estimate of phylogenetic relationships on all the
sequence evidence that is available, or, if on the basis of other evidence
we have reason to think that some of the data are likely to be misleading,
should those be eliminated from the analysis? If the latter, what are the
criteria for choosing that portion to be eliminated? Moreover, what jus-
tifications do we advance to establish the validity of those criteria? Do we
treat all characters (in this case, nucleotide positions) equally ("weighted
uniformly") or should some be weighted more than others (eliminating characters might also be viewed as a form of weighting)? Do we treat all character-state transformations at any given nucleotide position equally, or are they to be weighted differently? In both cases, how is weighting itself to be justified and how might weights be determined objectively? These questions can only be a prelude to a full discussion of how parsimony is to be applied inasmuch as many other issues (e.g., identifying signal, assessing topological reliability, the notion of congruence) also have a bearing on our choices regarding systematic procedure (Cedergren et al., 1988; Kluge, 1989).

One theme of this chapter has been to ask what might be the best approach for applying parsimony procedures to sequence data. Because the problem involves so many complex issues, not the least of which are the philosophical tendencies of the investigator, it is unlikely that the systematic community will agree on a unified answer. If one sees the goal of systematic research to be the discovery of the one true history of life, and accepts the precept that we will never have certain knowledge of that history, then our only recourse would seem to be to accept the hypothesis of relationships which is most consistently revealed by the data regardless of which analytical techniques might have been employed (this does not imply, of course, that all methods of analysis are necessarily equally useful or appropriate). This is essentially an appeal for acceptance of stability of cladistic signal and to identify as noisy those cladistic patterns that are ambiguous or less consistently revealed.

There has been substantial support in the literature for weighting among nucleotide positions (e.g., within codons) or among alternative character-state transformations within positions, and many weighting procedures have been suggested (e.g., Farris, 1969; Altschul et al., 1989; Williams and Fitch, 1989, 1990). Although there is a rationale for thinking that weighting will improve the resolution of cladistic signal, what is less clear is which method of weighting is "best" in any case. In principle, at least, a tree derived from weighted data should be a better or more reliable estimate of the true phylogeny than a tree derived from unweighted data, simply because the effect of weighting should be to emphasize the best available evidence (Carpenter, 1988). Yet, a final conclusion regarding the reliability of the resulting tree can only be reached by comparison to results based on more data and an assessment of congruence of their cladistic signal. It is possible, after all, that various assumptions about the evolutionary process that may underlie a particular method of weighting are themselves incorrect and thus may have led to spurious results. An unfortunate tendency within molecular systematics has been the application of a weighting scheme and the uncritical acceptance of the resulting tree. Currently available parsimony algorithms are designed to produce a single point solution; that is, a single most-parsimonious tree or a set of equally most parsimonious trees. Yet, it is difficult to judge from a single most-parsimonious tree whether a cladistic grouping represents a strong (stable) signal found in
the data or whether the data are noisy with respect to that group’s resolution (e.g., within the hominines). Thus, even if the data are weighted, questions about the reliability of the result still exist in addition to those raised by any justification of the weighting methods themselves.

The analyses undertaken in this study explore ways of revealing phylogenetic signal in sequence data through parsimony procedures. Taken together, the results of this analysis suggest that if there is any phylogenetic signal inherent in the data, parsimony methods will find it. The ultimate arbiter of this conclusion must be congruence with phylogenetic relationships inferred from other data sets, which in the case of the primates examined here support the conclusions of this chapter (e.g., Hayasaka et al., 1988b; Miyamoto and Goodman, 1990). Taken separately, each approach to parsimony reveals a significant portion of that signal, but resolution in some parts of the tree may remain ambiguous. Most-parsimonious trees based on the entire 898 bp are essentially congruent with the distance tree of Figure 10-1, there being two equally most-parsimonious trees differing only in the placement of chimpanzee with human or with gorilla; the single most-parsimonious tree using transversions alone resolved that ambiguity in favor of chimpanzee + human. An assessment of the strength of the phylogenetic signal is best described, however, by comparison of the majority-rule consensus trees derived from near-most-parsimonious solutions using global (Fig. 10-2) and transversion (Fig. 10-3) parsimony, along with those derived from bootstrap analysis of global (Fig. 10-5) and transversion (Fig. 10-6) parsimony solutions. Indeed, a comparison using these four methods of analysis demonstrates a strong phylogenetic signal in even the smallest subsets of the data, including the first 225 bp fragment and the 198 bp tRNA fragment. Caution must be exercised in generalizing these results inasmuch as this is only one data set, but parsimony analyses such as these appear to constitute a powerful methodological tool for resolving phylogenetic signal even in a small amount of data. Hayasaka et al. (1988a:637) concluded that the high rate of homoplasy in mtDNA makes the more distant relationships of the tree less reliable than those that are more recent, yet parsimony analysis revealed a strong phylogenetic signal across the entire tree.

Phylogenetic Inference Using Sequence Data: Some Comments on Alternative Methods

There has been considerable discussion in recent years about which method of phylogenetic inference is the most appropriate for DNA sequence data. Efforts have been made to adjudicate this issue by simulation experiments in which a data set is generated based on a “known true” phylogeny and given assumptions as to how sequence evolution proceeds (Tateno et al., 1982; Sourdis and Krimbas, 1987; Saitou and Imanishi, 1989). Different methods are then applied to the data set to see if they recover the “true
phylogeny." These simulations are perhaps useful for identifying conditions under which a given method may have limitations, but they are encumbered with philosophical and empirical problems. Because certain knowledge of phylogenetic relationships is beyond us, it is questionable whether some artificially generated "truth" can serve as the arbiter of the best method for recovering that which we cannot hope to know with certainty. The evolutionary models underlying these simulations also suffer from a critical methodological circularity: it is reasonable to assume that a detailed and empirical understanding of the mechanisms of molecular change will not be obtained unless it is based on prior hypotheses about the phylogenetic relationships of organisms. Consequently, to incorporate assumptions about the evolutionary process into a simulation procedure that will be used to choose a method of phylogenetic inference may bias our recovery of those relationships, which themselves will bias and confound our understanding of processes of molecular evolution. Perhaps as important, these simulation experiments generally take as their model of evolutionary change one that most closely applies to noncoding, randomly evolving DNA. In practice, however, the majority of molecular systematists are attempting to infer relationships from sequences that will seldom, if ever, evolve under the constraints specified by those models.

In real-world cases, we might expect that when there is a strong phylogenetic signal inherent in the data, virtually any method will recover it. When data are extremely noisy, say as a result of too little data being used to resolve very close internodal distances or of too much homoplasy, then perhaps no method would be expected to resolve a reliable signal. Somewhere in between these extremes we might expect that different methods will sometimes yield different phylogenetic hypotheses for a given set of data. Choice among those hypotheses should not, however, be based on an appeal to method, but on comparison and congruence with phylogenetic inferences derived from other data.

At this time, three primary methods of phylogenetic inference are applied to sequence data; namely, parsimony analysis, distance analysis, and maximum likelihood. Distance analysis has had both its defenders (Felsenstein, 1984; Nei, 1987) and critics (Farris, 1981, 1985). Tree-building algorithms that work on distance matrices are necessary for certain kinds of data, such as DNA-DNA hybridization or immunological distances, but it is questionable whether they should be applied to discrete character data such as sequences. At issue is not whether distance analysis of sequences is able to reveal phylogenetic structure, for clearly it can. More to the point is the inevitable loss of information in converting sequence data to distances (Farris, 1981, 1985; Penny, 1982), their sensitivity to saturation effects so that they must be corrected, the loss of flexibility to examine the deep structure of the data underlying the results, and the inability of most distance algorithms to examine the structure of the large suite of trees having near minimum-length fit. Taking the distance analysis of the primate data
set as an example, Hayasaka et al. (1988a) did not cluster using original distances but corrected them for multiple substitutions (Gojobori et al., 1982) and then clustered on those new distances. A single tree was produced by the neighbor-joining method (Saitou and Nei, 1987) (Fig. 10-1) but without an assessment of whether it was the best-fit tree. That tree, moreover, was not compared quantitatively or qualitatively to other near-minimum-length trees. Many different models have been proposed to "correct" for multiple substitutions. Yet, even though substitution probabilities can be estimated from observed nucleotide frequencies, there is no guarantee that those probabilities actually reflect the dynamics of processes that produced the observed nucleotide variation. That being the case, it remains to be seen to what extent these different models that are being applied to sequence data will influence inferred phylogenetic structure, not only of the minimum-length tree, but of those trees close to that solution.

Maximum-likelihood methods are more developed in theory than in practice (Felsenstein, 1981; Bishop and Friday, 1985; Saitou, 1988, 1990; Fukami and Tateno, 1989), and we will not discuss them extensively here. A choice among trees using maximum-likelihood methods depends upon accepting an underlying model of evolutionary change. Often, that model incorporates equal probabilities of substitutional change from one nucleotide to another, but some have been developed to account for differences in rate between transitions and transversions (Saitou, 1990). Calculation of likelihood functions for trees is computationally difficult, and the method cannot be extended easily beyond cases for five or six taxa. It seems that if more realistic, and thus more complicated, models of evolutionary change are to be incorporated, computational difficulty is sure to increase.

Cedergren et al. (1988:102) note that parsimony methods have virtue in phylogenetic inference using sequence data because they result "in the most economical reconstruction of mutational history, with no assumptions and with the minimum of coincidence and unobserved changes...." Although it is arguable whether parsimony is assumption free (all scientific methods designed to discover pattern involve some assumptions), the strengths of parsimony analysis over other methods of analyzing sequence data are easy to enumerate. With parsimony, it is possible to use a minimum of assumptions about the processes governing molecular evolution and still obtain interpretable results. Weighting of characters or character-states involves assumptions, but that too can be accomplished conservatively such that narrowly constrained assumptions about molecular processes are avoided. Today's parsimony algorithms are fast and efficient at searching among all possible trees, thus guaranteeing that not only will the most-parsimonious tree be found, but that the cladistic structure of near-most-parsimonious trees can be examined. Moreover, with parsimony it is possible to examine and evaluate the data supporting each cladistic component on a tree. For these reasons alone, parsimony is the preferred method for analyzing DNA sequences.
ACKNOWLEDGMENTS

We would like to thank T. E. Dowling, D. M. Hillis, T. Jones, D. Mindell, and especially M. M. Miyamoto for their helpful comments on the manuscript. We thank E. Chipouras and M. M. Miyamoto for providing the calculations for Table 10-3. This research was supported by the National Science Foundation through grants BSR-8805957 and BSR-9007652.

REFERENCES


