INVITED REVIEW

Phylogenetic Tree-building

DAVID A. MORRISON

Molecular Parasitology Unit, University of Technology Sydney, Westbourne Street, Gore Hill, NSW 2065, Australia

(Received 13 October 1995, accepted 3 March 1996)

Abstract—Morrison D. A. 1996. Phylogenetic tree-building. International Journal for Parasitology 26: 589-617. Cladistic analysis is an approach to phylogeny reconstruction that groups taxa in such a way that those with historically more-recent ancestors form groups nested within groups of taxa with more-distant ancestors. This nested set of taxa can be represented as a branching diagram or tree (a cladogram), which is an hypothesis of the evolutionary history of the taxa. The analysis is performed by searching for nested groups of shared derived character states. These shared derived character states define monophyletic groups of taxa (clades), which include all of the descendants of the most recent common ancestor. If all of the characters for a set of taxa are congruent, then reconstructing the phylogenetic tree is unproblematic. However, most real data sets contain incongruent characters, and consequently a wide range of tree-building methods has been developed. These methods differ in a variety of characteristics, and they may produce topologically distinct trees for a single data set. None of the currently-available methods are simultaneously efficient, powerful, consistent and robust, and thus there is no single ideal method. However, many of them appear to perform well under a wide range of conditions, with the exception of the UPGMA method and the Invariants method.

Key words: Phylogeny; evolution; cladistics; cladograms.

INTRODUCTION

If we accept the proposition that evolution exists, then meaningful comparisons among organisms must ultimately include a phylogenetic context (Maddison & Maddison, 1992). This is because the evolutionary relationships among a group of taxa constrain any other possible relationships that might exist. It is thus not surprising that in biology there has been a long history of attempts to deal with the reconstruction of genealogical history (Nelson & Platnick, 1981; Mayr, 1982; Stevens, 1994), notwithstanding the difficulties associated with producing testable hypotheses about historically unique events.

It is, however, only in the last 30 years that widespread attempts have been made by systematists to produce explicit and repeatable methods for the construction of phylogenetic trees (Felsenstein, 1982), notably with the translation from German into English of Hennig (1966). These methods are now usually referred to as “cladistics”, and the evolutionary diagrams they produce “cladograms”, to distinguish them from all prior phylogenetic studies, many of which were neither explicit nor repeatable (note that I am using the word cladistics to include a wide class of explicit and repeatable phylogenetic analyses, which may be a broader definition than would be accepted by many phylogeneticists, who would restrict the term to the more strictly Hennigian methods). The current interest in cladistic analyses has inevitably led to a proliferation of data analysis techniques, and the apparent plethora of methods for phylogenetic analysis is typical of a young science still coming to terms with both its aspirations and its constraints.

It is therefore important for practitioners to understand the limitations of the available techniques as well as to appreciate their capabilities (Felsenstein, 1982). Unfortunately, the current wide choice among possible phylogenetic methods seems to be daunting.
for many people, and they thus acquire little knowledge of the relative advantages and disadvantages of the various methods. This is unfortunate, because the choice of data-analysis methods should be based on their apparent appropriateness for the data at hand, rather than on the local availability of computer programs or on historical inertia (Hillis, Allard & Miyamoto, 1993). It is my purpose here to review the phylogenetic inference methods that are currently available, and to indicate what is currently known about their strengths and weaknesses. This will allow a more informed decision to be made when assessing which of the methods might be appropriate for a particular data set.

Along the way, I will attempt to make clear some of the aspects of phylogenetic analysis that are obviously misunderstood by non-specialists, and to dispel a few widely-held misconceptions. My discussion will focus on molecular sequence data (particularly DNA and RNA), since trees derived from this source are increasingly those with which parasitologists are working (e.g., Nadler, 1990), there being a limit to the usefulness of phenotypic characteristics for constructing phylogenetic trees for most parasites. Indeed, it is the proliferation of molecular data sets that is introducing many people from outside of systematics to the science of phylogeny (Miyamoto & Cracraft, 1991; Hillis, Huelsenbeck & Cunningham, 1994a); and it is for this reason that phylogenetic analyses were not necessarily based on non-experts. Furthermore, many of the recent advances in cladistics have been motivated by attempts to deal with problems that are specific to molecular data (Penny et al., 1990), and the similarities and differences between traditional and molecular phylogeny thus need to be emphasized.

I do not intend to describe the tree-building techniques in any great detail, partly because most of them are covered in the excellent review by Swofford & Olsen (1990) and in the books by Nei (1987) and Li & Graur (1991), and partly because nothing is more intimidating to most biologists than mathematics. I concentrate much more on those aspects of phylogenetic analysis that are likely to be of most practical and theoretical interest to non-experts. I start by summarizing the principles of cladistic analysis, before proceeding to a review of the alternative methods for constructing cladograms. A number of introductory reviews covering similar topics, but from different perspectives, include those of Felsenstein (1988), Olsen (1988), Sneath (1989), Beanland & Howe (1992), Hillis et al. (1993) and Stewart (1993); and there is also the worthwhile introductory book by Forey et al. (1992).

**CLADISTICS**

The study of evolutionary processes has often been considered to be unscientific because it deals with historically unique events (Popper, 1957). Hypotheses concerning these events are thus not universal (in either space or time) and, therefore, they are considered to be untestable in the contemporary world. The sociological development of phylogenetic analysis has consequently been based largely on the creation of what have been called “evolutionary scenarios” describing the presumed genealogical history of the organisms under study. The number of such scenarios that may be created is, of course, limited solely by the imagination of the researcher, and none of the scenarios are likely to be open to falsification. Cladistic analysis can thus be seen as an attempt to base phylogenetic analysis on a more objective footing, where the phylogenetic hypotheses are explicitly stated, along with the evidence supporting (and contradicting) them, and are then subjected to quantitative testing. Its practitioners therefore claim that cladistics is designed to make phylogenetic analysis into an hypothetico-deductive science, where explicit hypotheses are subjected to repeatable attempts at falsification.

Note that the claimed advantages of cladistic analysis are not intended to denigrate pre-cladistic biologists, nor is there any suggestion that these biologists did not apply their minds to phylogenetic questions. However, it is clear that pre-cladistic phylogenetic analyses were not necessarily based on repeatable methods that produced explicit hypotheses of evolutionary relationship which could be subjected to falsification (Nelson & Platnick, 1981). Furthermore, the taxonomic groups produced by pre-cladistic biologists were not necessarily monophyletic (see below), and therefore did not always reflect evolutionary history. Post-Darwinian biologists have had the unenviable task of producing taxonomic schemes that should, in theory, reflect evolutionary history, without any theoretical framework for how they should go about discovering what the evolutionary history actually was (Stevens, 1994). Cladistics is an attempt to provide this theoretical framework.

Cladistic analysis as an approach to phylogeny attempts to group taxa on the basis of their ancestry. In the analysis, taxa are grouped in such a way that those with historically more-recent ancestors form groups nested within groups of taxa with more-distant ancestors. The analytical technique is based on a widely-held view of the mode of the evolutionary process: species are lineages undergoing
divergent evolution with modification of their 
intrinsic attributes, the attributes being transformed 
through time from ancestral to derived states. Thus, 
if a species is a group of populations, then if 1 group 
of populations acquires a new (derived, advanced or 
apomorphic) character state while the rest do not (i.e. 
they retain the ancestral, primitive or plesiomorphic 
state) then these populations constitute a new species. 
This new species then forms a separate historical 
lineage that diverges from the other populations, and 
maintains its own historical tendencies and fate. 
Evolution can thus be thought of as a branching 
sequence of historical lineages, and indeed the only 
diagram in Darwin (1859) represents precisely such a 
branching sequence.

When a character changes from an ancestral state 
to a derived state in a lineage it is historically unique 
(a novelty), and it will be passed on to all of the 
descendants of that lineage (even if the character is 
later modified into something else). Therefore, the 
branching sequence of evolution can be deduced by 
searching for nested groups of shared derived char-
acter states (synapomorphies) among the taxa being 
analysed. So, if a derived character state is observed 
in 2 or more taxa, then we can hypothesize that they 
share this apomorphy because they are descended 
from a common ancestor, and that they inherited the 
apomorphy from that ancestor. The possession of a 
shared ancestral state (a symplesiomorphy) tells us 
nothing about the phylogeny of the taxa, since this 
state was inherited from an ancestor that is also held 
in common with those taxa possessing the derived 
state. Thus, cladistic analysis is simply the search for 
nested sets (a hierarchy) of synapomorphies among 
the taxa. Each synapomorphy represents an ancestral 
lineage that has diverged from its related lineages, 
thus being contemporary evidence for a prior 
evolutionary event (Table 1). Clearly, the only 
characters that are of use for a cladistic analysis of a 
group of contemporary taxa are those features that 
reflect their evolutionary history. There is, of course, 
no simple method for determining which features 
these are, but at least the method forces the practi-
citioner to be explicit about the characters that have 
been chosen for the analysis.

Homology

For this analytical technique to work, homologous 
rather than analogous character states must be com-
pared across the taxa (Hall, 1994). That is, for all of 
the taxa we must compare like with like, particularly 
with reference to the evolutionary origin of the 
attributes. Characters and their states are thus hy-
potheses of evolutionary homology. As an example, 
for Character 16 in Table 1 possession of denticles, 
dermal scales, epidermal scales, feathers and hair are 
all considered to be homologous character states of a 
single character. In practice, characters and their 
states are postulated as homologous on the basis of 
their structural, positional, ontogenetic, composi-
tional and/or functional correspondences; and they 
are postulated between different taxa so as to maxi-
mize the number of one-to-one correspondences of 
their parts (Stevens, 1984). The choice of characters 
to be included in a cladistic analysis may be some 
what arbitrary, but can include intrinsic attributes 
such as morphology, anatomy, embryology, behav-
ior, physiology, ultrastructure, cytology, biochemis-
try, immunology, ecology and even geography. The 
important points are that the characters used in the 
analysis are hypothesized to reflect the evolutionary 
history of the taxa, and that the character states of a 
single character are hypothesized to have a unique 
evolutionary origin.

Concepts of homology are often intuitively obvi-
ous when dealing with, for example, morphological 
data, and these concepts can be put into practice 
through a detailed study of ontogeny (e.g., the 
homologies postulated in Table 1 are thus mostly 
unproblematic), and the alignment of molecular 
sequences is the direct equivalent of these homology 
assessments (Mindell, 1991). The concepts of hom-
ology in molecular and morphological studies are 
thus fundamentally the same (de Pinna, 1991; 
Williams, 1993). So, alignment of molecular 
sequences involves a series of hypotheses of hom-
ology among the taxa, with 1 hypothesis of homology 
for each position (nucleotide or amino acid) in the 
sequence. It is important to recognize this, because 
it is clear that often very little attention is paid 
by molecular biologists to this point — to a 
morphologist the assessment of homology is an 
important (and time-consuming) component of phy-
logenetic analysis, but to a molecular biologist the 
assessment of homology is apparently often an after-
thought. The correct formulation of hypotheses of 
homology is just as important for molecular data as it 
is for morphological data; and it is clear from the 
literature that many of the so-called controversies 
about the phylogeny of particular groups of para-
sites are based as much on disagreements about 
sequence alignment as on disagreements about actual 
evolutionary events (e.g., Ellis & Morrison, 1995).

Unfortunately, for molecular data there is 
little possibility of further investigations (such as 
ontogeny) to assess homology, and so in practice 
homology assessment is very different for molecule 
studies (Mindell, 1991). Positional homology can 
be represented by either identical character states
Table I—Some phenotypic data for extant vertebrates that might be useful for reconstructing their evolutionary history

<table>
<thead>
<tr>
<th>Character</th>
<th>Lampreys</th>
<th>Sharks</th>
<th>Teleosts</th>
<th>Lungfish</th>
<th>Frogs</th>
<th>Salamanders</th>
<th>Turtles</th>
<th>Lizards</th>
<th>Snakes</th>
<th>Crocodiles</th>
<th>Birds</th>
<th>Mammals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Internal skeleton</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2 Jaws</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>3 Ossified skeleton</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>4 Internal nostrils</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>5 Atrial septum</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>6 Four limbs</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>7 Teeth</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
</tr>
<tr>
<td>8 Amniotic egg</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
</tr>
<tr>
<td>9 Temporal fenestrae</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
<td>two</td>
</tr>
<tr>
<td>10 Hemipenes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>11 Suspensorium streptosyous</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12 Antorbital fenestrae</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>13 Lateral fenestrae ossified</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>14 Gizzard</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>15 Homeothermy</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>16 Body covering</td>
<td>scale-less</td>
<td>denticles</td>
<td>scales</td>
<td>scales</td>
<td>endermis</td>
<td>epidermis</td>
<td>scales</td>
<td>scales</td>
<td>scales</td>
<td>scales</td>
<td>scales</td>
<td>scales</td>
</tr>
</tbody>
</table>

These 16 characters each have at least 2 character states that are hypothesized to be homologous, one of which is postulated to be ancestral (the states “no”, “none”, and “scale-less”). The other (one or more) states are postulated to be derived from this ancestral state, and the taxa that share these derived states are then hypothesized to have inherited this state from a common ancestor. These shared derived character states (synapomorphies) thus group the taxa into sets that reflect their evolutionary history. Note that Character 1 has only 1 state in this data set, because the derived state has been inherited by all of these taxa, and it thus forms a synapomorphy for the group (vertebrates) as a whole. Furthermore, Characters 9 and 16 have more than 2 homologous states, thus implying that the ancestral state has been modified more than once during the evolutionary history of these taxa, each of these derived states forms its own set of taxa. A possible cladogram derived from these data is shown in Fig. 1. The data are from Romer (1971) and Maddison & Maddison (1992). Note that this non-parasitological example has been chosen because the homology assessments and the resulting cladogram are not controversial, which is something that cannot be said for any parasitological example known to me.
mutations), or insertions/deletions (indels) in 1 or more sequences. The most problematic aspect of sequence alignment is the positioning of indels, and this problem becomes more acute for more divergent sequences. It is worthwhile in this context to distinguish between gaps, which are introduced into the sequences by the alignment procedure, and indels, which are actual mutation events (Olsen, 1988); clearly, the objective is to introduce into the sequences only gaps that truly represent indels.

There are 3 possible scenarios for the degree of difficulty of sequence alignment. Firstly, there may be relatively few indels, in which case a robust sequence alignment can usually be produced by hand. Such a situation is shown, for example, by protein-coding parts of mtDNA (Miyamoto & Cracraft, 1991) and the 5.8S rRNA gene (Chase et al., 1993). Secondly, the sequence may represent a molecule for which there is an a priori biological model of secondary structure in which certain active sites must be maintained; the alignment is then constrained by the base-pairing of the model (Olsen, 1988). Such a situation is shown, for example, by the 5S (Specht, Wolters & Erdman, 1990), 12S (Gutell et al., 1985) and 18S (Van de Peer et al., 1994) rRNA genes.

Thirdly, there may be many indels and no a priori structure model. Under these circumstances it is usual to use a mathematical algorithm to produce the alignment. These algorithms all attempt to produce a sequence alignment that optimizes some chosen criterion of match between the individual sequences (cost). That is, the sequences are compared using a pattern-matching process that searches for correspondence between the elements of the sequences, introducing gaps into the sequences as required to maximize some criterion for optimality of the correspondence. There are many algorithms currently available (see Waterman, 1989; Doolittle, 1990; Chan, Wong & Chiu, 1992; and McClure, Vasi & Fitch, 1994), which optimize a variety of mathematical functions measuring the overall alignment cost. When there are more than 2 sequences, most of these algorithms use exact procedures (see below for a definition) to align the sequences pair-wise, but then use heuristic procedures (see below for a definition) to braid these alignments into a multiple alignment. Thus, these procedures do not guarantee to produce the globally optimal alignment; nor do they guarantee that the optimal alignment (even if they could find it) represents the true alignment (Thorne & Kishino, 1992).

I am not going to discuss these various algorithms here, because this is an area of active research and very little is known about their theoretical or practical limitations (Thorne & Kishino, 1992; McClure et al., 1994). However, 1 important generalization may be made: the differences between the alignments produced by the various algorithms is often less than the differences produced by varying the "gap weights". These weights refer to the relative cost of inserting a new gap into a sequence or extending an already-existing gap, and there is no way of determining analytically what these weights should be (Rinsma-Melchert, 1993). Most of the computer programs that implement the alignment algorithms have default values for the weights that are designed to produce "biologically interesting" results, and very few molecular biologists seem to be willing to deviate from these default choices. It is, however, clear that to simply report that a particular computer program was used to align the sequences is meaningless (since the work cannot be verified) unless the weight values are also reported (Wheeler, 1995).

In dealing with the problematic nature of sequence alignment, molecular biologists often delete parts of their sequences from the cladistic analysis. The rationale for this is that those parts of the sequences that cannot be aligned reliably should be excluded from the estimation of the phylogeny (Olsen, 1988). This is probably very sensible (Smith, 1994), but unfortunately there is often no objective criterion given for deciding which parts of the alignment are ambiguous, the decisions usually being made by "visual inspection". Gatesy, DeSalle & Wheeler (1993) have suggested that those parts of the alignment that are sensitive to the gap weights (i.e. where the alignment varies significantly when the gap weights are changed) may constitute unreliable hypotheses of homology and may therefore be candidates for exclusion; and Ellis & Morrison (1995) have shown that for some organisms it is the double-stranded parts of rRNA that may contain most of the phylogenetic information when the sequences are aligned according to secondary structure. Thus, there is considerable room for further research into the problems of sequence alignment.

It is also important to recognize that there is also a more general level of homology assessment for molecular data. The sequences being compared must themselves be homologous rather than analogous (Fitch, 1970). Thus, only orthologous sequences will reflect the historical relationships of species, while paralogous sequences (e.g., the 2 sequences that result from a gene duplication) will reflect only gene history, xenologous sequences (e.g., recently-incorporated sequences such as result from horizontal gene transfer) will only partly reflect gene history, and plerology (e.g., the inter-mixture of exons and
introns) will only reconstruct a composite gene history (Patterson, 1988; Williams, 1993). All molecular studies thus rely on the assumption that the sequences from the taxa being compared are orthologous, and this may be a dubious assumption for distantly related taxa (Sneath, 1989). Any sequence for which orthology has not been established should be omitted from the analysis (Olsen, 1988).

Furthermore, for orthologous sequences there is the implicit assumption that the sequences being compared are actually from the organisms being studied, rather than from some other co-habiting organism. For example, it is now recognized for the ITS gene that many of the published sequences purporting to be from conifers and ferns are actually those of fungi, while the published Mimulus (monkey-flower) sequences are actually those of green algae (P. Weston, personal communication).

As a final observation on molecular data, it is worth emphasizing the distinction between species trees and gene trees. The result of cladistic analysis of molecular data is a gene tree (provided that the entire gene has been sequenced), hypothesizing relationships among the genes or genomes that have been sampled, whereas a species tree reflects the actual evolutionary pathways (Famili & Nei, 1988). The gene tree may be fundamentally incongruent with the true species phylogeny, with the genome tree, and with other gene trees (e.g., Cao et al., 1994; Cummings, Otto & Wakeley, 1995), due to various phenomena such as allelic polymorphism, introgression, lineage sorting, unequal rates of speciation and gene mutation, lateral transfer, hybridization, or mistaken orthology (Famili & Nei, 1988; Penny et al., 1990; Doyle, 1992; de Queiroz, Donoghue & Kim, 1995). Thus, for the reconstruction of phylogenetic history, a single gene may be, in practice, no more useful than a single morphological character (Doyle, 1992). It may, therefore, be wise to assume tacitly in molecular studies that the number of characters being used to reconstruct a phylogeny is equivalent to the number of positions in the sequence. Furthermore, a living organism is an integrated functioning whole, not just a collection of unrelated genetic attributes. Thus, an organism is a collection of interactions between genes, and between genes and their environment (i.e. a phenotypic whole), and it is the organism as a whole that takes part in the evolutionary process. Consequently, there is no more reason for genetics to reflect phylogeny than for anything else to do so (de Queiroz et al., 1995). In fact, morphological characters may be a better reflection, because they integrate many genetic and phenotypic characters.

Polarity

Having determined the homology of the character states, the key to cladistic analysis is the distinction between derived character states and ancestral states (character polarity). It is important to note that this is a local concept that applies only to a particular set of taxa. By this I mean that a character state is only considered to be derived relative to a specified ancestral state, and it may well be the ancestral state for a further derived state. As an example, for Character 16 in Table 1 possession of epidermal scales is a derived state relative to possession of dermal scales, but is an ancestral state relative to possession of either feathers or hair. Thus, possession of epidermal scales is a synapomorphy for Turtles+Lizards+Snakes+Crocodiles; Birds+Mammals, while possession of feathers is a synapomorphy for Birds and possession of hair is a synapomorphy for Mammals; this character does not supply a synapomorphy for the grouping only of Turtles+Lizards+Snakes+Crocodiles. Possession of epidermal scales is thus a synapomorphy at a more general level than is possession of either feathers or hair. There is thus a hierarchical relationship between ancestral and derived character states, and recognizing synapomorphies therefore involves determining the correct level of generality of homologies.

Clearly then, the success of a cladistic analysis rests on the correct determination of the relative apomorphy of the character states, and numerous criteria have been proposed for doing this (Crisci & Stuessy, 1980; Stevens, 1980; Bryant, 1992; Nixon & Carpenter, 1993). Most of these criteria rely on illogical arguments or on assumptions that are either false or untestable; and it is worthwhile recognizing only 2 objective possibilities: the direct method; and outgroup analysis. These are complementary methods, and both may thus be used in any 1 data set. For the direct method, the hierarchical relationship between ancestral and derived character states is observed directly, and it does not require any pre-existing hypotheses of character polarity. On the other hand, the outgroup analysis method does not directly observe character polarity, and it relies upon an hypothesis of the relationship of the taxa under study to their near relatives.

The direct method (Weston, 1988) states that: if 1 character state is possessed by all of the taxa that also possess the alternative state, and in addition it is possessed by some taxa that do not possess the other state, then it is postulated to be the ancestral state. For example, open gill slits are possessed by all chordates in at least the embryonic stage, but in tetrapod chordates these close early in development;
Fig. 1. A phylogenetic tree derived from the character data shown in Table 1. Almost all of the characters shown in Table 1 have derived states that are congruent with this cladogram; that is, the synapomorphies form a perfect series of nested sets. For example, Characters 10 and 11 form a set consisting of the Lizards+Snakes, while Characters 12-14 form a set of the Birds+Crocodiles; these characters are plotted on the branch referring to the hypothesized ancestor where the derived character state arose (and the other characters can be plotted in similar fashion). The only exception to this congruence concerns Character 15, for which 2 evolutionary origins of the derived state must be postulated (i.e. it appears twice on the cladogram); that is, there is an apparent convergence on this cladogram, where the derived state is postulated to have arisen in 2 separate ancestors. This implies a mistaken hypothesis of homology (a homoplasy) for homeothermy in Birds and Mammals. Other cladograms could be constructed in which this convergence does not occur (i.e. in which Character 15 is a synapomorphy uniting Birds+Mammals), but in all of these cases at least 2 of the other characters must then be homoplasious.

thus, all chordates possess open gill slits but only some possess both open gill slits (early in development) and closed gill slits (later in development). Consequently, possession of closed gill slits is hypothesized to be derived relative to possession of open gill slits. This type of argumentation can be applied to many types of characters (Nelson, 1978; Weston, 1988, 1994), but it is probably of limited utility for molecular data. However, Weston (1994) has successfully applied the direct method to gene duplications (paralogy) to polarize the a and b sub-units of ATPase based on taxa from the archaeabacteria, eubacteria and eukaryotes, thus providing a "root" for the tree of life.

The outgroup analysis (or indirect) method (Wattous & Wheeler, 1981) states that: if a character state is found in both the ingroup (the group of taxa under study) and also in the outgroup (the sister group of taxa), then it is postulated to be the ancestral state. For this method to provide unequivocal evidence of character polarity, the outgroup should consist of at least 2 sequential sister groups (Maddison, Donoghue & Maddison, 1984). For example in Fig. 1, if we were interested in the phylogeny of the tetrapods (the ingroup thus consisting of Frogs+Salamanders+Turtles+Lizards+Snakes+Crocodiles+Birds+Mammals), then the relevant sister groups (the outgroup) would be at least the Lungfish and the Teleosts. Thus for Character 8 in Table 1, possession of an amniotic egg is hypothesized to be apomorphous relative to the lack of such an egg because all members of the outgroup and some members of the ingroup lack it, while only some members of the ingroup possess it. This type of argumentation relies on the existence of a corroborated higher level phylogeny for the taxa being studied, because we need a priori knowledge of the sister groups of the ingroup. Such higher level phylogenies in turn may rely on other outgroup comparisons, and so on in a regress back to the origin of life. Ultimately, we must rely on the direct method for at least some of the characters in some of the analyses.

Outgroup analysis is the most common type of argumentation in cladistic analysis, but it is now usually implemented in a different way from its original formulation, particularly when analysing molecular data (Smith, 1994). Instead of first determining character state polarity and then producing the cladogram, it is now more common to produce an
unrooted cladogram (a network) based on simultaneous analysis of all of the characters of both the ingroup and the outgroup and then to determine the root of the tree using the position where the outgroup joins the ingroup (see below). This type of method was first described by Farris (1972), and Nixon & Carpenter (1993) provide a recent detailed summary.

In this method, there are no guarantees that any particular number of outgroups or any particular choice of outgroups will ensure that the cladogram accurately reflects the evolutionary history (Li et al., 1987; Smith, 1994; Adachi & Hasegawa, 1995). In practice, the phylogenetic inferences will be more robust if more outgroup taxa are chosen rather than fewer, and the more closely-related these taxa are to the ingroup (Hendy & Penny, 1989; Wheeler, 1990; Nixon & Carpenter, 1993). For molecular data, a distantly-related outgroup may have surpassed the point of saturation of base substitutions, and there will thus be a loss of phylogenetic signal through evolutionary time as a result of random sequence noise (Smith, 1994). The outgroup may thus, in effect, be random, and as such it will simply join the rest of the cladogram on the longest internal branch in the ingroup and will itself have a long terminal branch (Wheeler, 1990). The best strategy may be for the outgroup to consist of several taxa from the sister group to the ingroup (Smith, 1994).

Conclusion

Ideally, all hypotheses of homology and relative apomorphy will be congruent with one another. That is, the sets of synapomorphies for a group of taxa will form a perfect nested hierarchy, and the polarization of the character states provides a root for the tree. The construction of a cladogram from the data would then be unproblematic (Fig. 1); and it was indeed under these circumstances that Hennig (1966) introduced his phylogenetic method. In most of the real world, there are, however, always characters that are incongruent with each other (homoplasies). These incongruent characters are postulated to result from either reversals, where a derived character state reverts to the ancestral state, parallelisms, where the same derived character state arises in separate evolutionary lineages, or convergences, where superficially similar character states have arisen in separate lineages. Homoplasies are thus mistaken hypotheses of homology.

In order to deal with the common existence of homoplasies, a range of tree-building techniques has been developed in cladistics. Each technique implements a different stratagem for how homoplasies are to be treated, and each technique may thus produce a different hypothesis of phylogenetic history for the taxa being analysed. If a data set is perfectly congruent (i.e. all of the characters reflect the same speciation and phylesis events), then all of the techniques will produce the same cladogram, which will be the cladogram produced by the original method of Hennig (1966). However, apparent incongruences almost always exist in the real world, and therefore a range of methods has been developed to try to detect the phylogenetic pattern that underlies the apparent contradictions.

Cladistics is basically axiomatic, in the sense that if the assumptions (the axioms) are accepted then the rest (the corollaries) follows directly from them. Thus, like all axiomatic propositions, if one or more of the assumptions are invalid, then the rest of the edifice falls with them. The key assumption is that evolution is mainly a divergent process. Homoplasies are thus assumed to represent "errors", rather than evidence in favor of some form of alternative evolutionary process (i.e. that does not produce a tree-like set of relationships), such as hybridization, endosymbiosis, recombination, gene duplication (producing pseudogenes), or lateral transfer. You might like to keep this in mind when you next attempt to interpret an apparently fully-bifurcating cladogram, as such alternative evolutionary processes are increasingly being recognized as relatively common.

Phylogenetic trees

For the purposes of cladistics, it is assumed that the phylogenetic pattern of evolutionary history can be represented as a branching diagram like a tree (a cladogram), with the terminal branches (or leaves) linking the taxa being analysed and the internal branches (or internodes) linking hypothesized ancestral taxa (nodes) (Fig. 1). The cladogram is thus usually constructed purely from a knowledge of contemporary taxa, but it implies the existence of ancestors with particular characteristics. Traditionally, the descendants of an ancestor on a cladogram are called daughters, while the siblings after a speciation event are called sisters (so a descendant is a daughter relative to its ancestor and a sister relative to its other sibling). Note that if either of the daughters undergoes further speciation then the sister to a particular contemporary taxon may actually be a group of contemporary taxa (i.e. all of the descendants of its sister), thus the sister to the Turtles in Fig. 1 is the group formed by the Lizards+Snakes+Crocodiles+Birds+Mammals.

Two types of evolutionary information can be represented on a cladogram. Firstly, the speciation
events in evolutionary history are represented by the branching sequence of the tree. Secondly, the phylesis events may be represented by the relative length of the branches on the tree (in which case it is sometimes referred to as a phylogram). Thus there is an explicit distinction between those evolutionary events that modify a taxon without causing new taxa to originate (phylesis) and those evolutionary events that do lead to the origin of new taxa (speciation). Both of these events are involved in evolutionary divergence among taxa. However, only the branching sequence (the topology) is essential to a cladogram, and indeed this is the only information shown in Fig. 1; the equivalent tree showing both speciation and phylesis is shown in Fig. 2. This is not to say that the phylesis events are not of interest, but merely that their pattern is not an essential part of the interpretation of the tree; and indeed only some of the phylesis can be shown on the tree (i.e. only for those characters being used) while all of the speciation may be included. Interestingly, the book of Darwin (1859) is basically about phylesis (the transformation of 1 species into another) rather than about speciation (the formation of 2 daughter species from 1 ancestral species), in spite of its title.

The branching sequence of a cladogram organizes taxa into monophyletic groups (clades), and the best way to discuss the evolutionary pattern represented by a cladogram is in terms of which taxa form monophyletic groups on the tree. A monophyletic group is: a group of taxa descended from a single ancestor and which includes all of the taxa descended from this ancestor. Groups that do not include all of the descendants of their most recent common ancestor are called paraphyletic groups. Thus in Fig. 1 Crocodiles+Birds form a monophyletic group, as do Lizards+Snakes+Crocodiles+Birds, and also Lizards+Snakes+Crocodiles+Birds+Mammals, but Lizards+Snakes+Crocodiles form a paraphyletic group (because their most recent common ancestor also has Birds as a descendant); the Class Reptilia as usually defined is thus not a monophyletic group (it would have to include all of the members of the traditional Class Aves and Class Mammalia as well).

Simple though it may seem, it is my impression that the recognition of monophyletic groups, and thus the description and interpretation of the evolutionary pattern shown by a cladogram, appears to be one of the more difficult aspects of cladistics for non-experts (see also the specific examples discussed by Penny et al., 1990 and Brooks & McLennan, 1992).

It should also be recognized that any 1 cladogram can actually be drawn as a tree in many different ways (Fig. 3). This is because the cladogram only specifies the temporal order of the branching
sequence from the root to the tips, and which
dughter is drawn on the tree to the left of the
ancestor and which is drawn to the right has no
evolutionary meaning. Consequently, the taxa can be
rotated around each internal branch on the tree
without in any way affecting the meaning of the
cladogram. The best analogy here is to think of a
mobile hanging from the ceiling, and to think about
the way in which its parts can rotate freely. The
mobile does not change its basic form no matter how
much the parts are rotated, and similarly the phylo-
genetic information in the cladogram does not
change no matter which way we rotate the branches.
It is thus very easy for non-experts to think that 2
trees are different, when in fact the cladogram still
specifies exactly the same evolutionary relationship
among the taxa (Fig. 3).

The distinction between rooted and unrooted trees
is very important (Fig. 4), but it is also often prob-
lematical for non-experts. As noted above, the most
common method of cladistic analysis (particularly for
sequence data) involves producing an unrooted tree
and then rooting the tree using the outgroup. This
rooting of the tree then indicates the direction of
evolutionary change on the cladogram; and this
allows hypotheses of relative character-state polarity
(ancestral vs derived) to be produced, and also allows
inferences about the composition of monophyletic
groups to made. So, unrooted trees do not indicate
character polarity, nor do they unequivocally show
monophyletic groups. This is because a group that
appears to be monophyletic on an unrooted tree will
be paraphyletic if the tree is rooted within this group.
For example, in Fig. 4 the group Taxon1+Taxon2+
Taxon3 is only indicated as monophyletic in Tree E,
because this is the only tree of those shown in which
the root lies outside of this group. Any specified
unrooted tree can actually be rooted on any of
its branches (Fig. 4), which is why it is important to
have an objective criterion (such as an outgroup, or
some characters polarized by the direct method) for
determining the position of the root.
Fig. 4. Several of the rooted phylogenetic trees that can be derived from a single unrooted phylogenetic tree. (A) Shows an unrooted cladogram of the hypothetical relationships among 5 taxa; (B) is the rooted cladogram that results from rooting the tree in (A) on the terminal branch leading to Taxon1; (C) is the rooted cladogram that results from rooting the tree in (A) on the terminal branch leading to Taxon2; (D) is the rooted cladogram that results from rooting the tree in (A) on the terminal branch leading to Taxon3; (E) is the rooted cladogram that results from rooting the tree in (A) on the terminal branch leading to Taxon4. The unrooted tree in (A) could also be rooted on the terminal branch leading to Taxon5. The unrooted tree in (A) could also be rooted on the terminal branch leading to Taxon2, on the terminal branch leading to Taxon4, and on the internal branch connecting Taxon4+Taxon5 with the other taxa. There are thus as many possible rooting points on an unrooted tree as there are branches on the tree. Note that the unrooted tree does not specify the direction of evolutionary change, while each of the rooted trees specifies a different direction of evolutionary change.

It is worthwhile pointing out here that, in spite of its limitations, an unrooted tree is, in itself, still a valuable phylogenetic commodity. This is because it limits the possible cladograms (rooted trees) that are supported by the data. For example, for 5 taxa there are 105 possible rooted trees, and yet the unrooted tree shown in Fig. 4 supports only 7 of these possible trees (i.e. we have rejected 98 possibilities). Furthermore, it will be possible to answer many phylogenetic questions using only the unrooted tree. For example, none of the rooted trees that can be derived from the unrooted tree shown in Fig. 4 has Taxon1+Taxon3 as a monophyletic group (since Taxon2 must also be included). Therefore, if this grouping was one of the hypotheses of interest, then we can test it adequately without ever needing to
Table 2—Some characteristics of commonly-used tree-building methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Procedure</th>
<th>Data</th>
<th>Taxa</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Unweighted pair-group</td>
<td>alg</td>
<td>dist</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>2 Neighbor-joining</td>
<td>alg</td>
<td>dist</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>3 Neighborliness</td>
<td>alg</td>
<td>dist</td>
<td>quart</td>
<td>morph</td>
</tr>
<tr>
<td>4 Minimum-evolution</td>
<td>opt</td>
<td>dist</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>5 Distance-Wagner</td>
<td>opt</td>
<td>dist</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>6 Least-squares</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>7 Maximum-parsimony</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>8 Weighted-parsimony</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>9 Compatibility</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>morph</td>
</tr>
<tr>
<td>10 Maximum-likelihood</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>molec</td>
</tr>
<tr>
<td>11 Invariants</td>
<td>opt</td>
<td>orig</td>
<td>quart</td>
<td>molec</td>
</tr>
<tr>
<td>12 Spectral analysis</td>
<td>opt</td>
<td>orig</td>
<td>all</td>
<td>molec</td>
</tr>
</tbody>
</table>

*Method chooses a tree based on an optimality criterion (opt) (i.e. a search method) or simply uses an algorithm (alg) (i.e. a constructive method).
*Method analyses the original data (orig) or data that have been converted to distances between taxa (dist).
*Method performs calculations on all taxa (all) or only on subsets of taxa (quart).
*Method suitable for both morphological and molecular studies (morph) or only for molecular studies (molec).

determine the root of the tree. Consequently, the production of an unrooted tree is a very big step towards the goal of producing a cladogram.

Finally, it is worthwhile remembering that a cladogram represents a set of hypotheses concerning evolutionary history. There are hypotheses about: character homology; character polarity; and phylogenetic relationships of taxa. All of these hypotheses are stated explicitly by the cladogram itself, irrespective of whether they have actually been shown on the diagram. The cladogram does not constitute an experimental test of the evolutionary history of the taxa, but it is a test of hypotheses of possible phylogenetic patterns. Because these hypotheses of pattern are stated explicitly, they are open to objective tests of their robustness. Furthermore, the hypotheses of pattern are open to further objective testing through cladistic analyses of other data sets for the same set of taxa. This may be as close as we will ever get to a hypothetico-deductive science of phylogeny.

Methods for constructing trees

Tree building. The tree-building methods that have been developed vary in many properties (Swofford & Olsen, 1990; Penny, Hendy & Steel, 1992). Some methods choose the tree from among all of those possible that either maximizes or minimizes some optimality criterion (search methods), while other methods merely follow an algorithm (a pre-defined sequence of operations) to produce a single tree (constructive methods). Some of the methods analyze the data matrix directly, while others require that the data have been converted to distances between taxa before the analysis (i.e. a matrix of all possible pair-wise distances between the taxa is calculated). Furthermore, some of the methods that use an optimality criterion evaluate that criterion over all taxa, while some methods evaluate the criterion on subsets of the taxa (usually 4 taxa, called quartets). Some methods have been developed specifically for sequence data, while others may be applied to any type of data. The more commonly-used methods (Huelsenbeck, 1995a) and their characteristics are summarized in Table 2, and are discussed below.

For those methods that have an optimality criterion, it is important to recognize that they are based on a double-level optimization, where first the optimality criterion is optimized for a given tree topology, and then the tree that optimizes these optimal values is selected (i.e. first you make each tree as good as you can, and then you find the tree that is the best). Methods thus need to be known that accomplish the optimization for a specified tree, and then methods need to be known for searching among the trees for the optimal one. The first class of methods will be unique for each tree-building procedure, but the second class of methods is more general.

There are several strategies for finding the optimal tree (or trees). Exact methods are those that guarantee to find the optimal solution, while heuristic methods are computationally-efficient strategies that should produce a solution that is at least close to the
optimal one even if it does not find the optimum. The most efficient (see below) of the exact methods involve the use of the branch-and-bound strategy (Hendy & Penny, 1982). The need for heuristic methods is simply one of practicality — the number of possible trees to be tested by the exact methods increases exponentially with increasing numbers of taxa (Felsenstein, 1978a). Heuristic search strategies first try to find a tree that is a good estimate of the final solution, usually by sequentially adding the taxa to the growing tree, adding each taxon in the optimal place on the tree. Most heuristic strategies then search through those trees that are similar to this initial tree in order to find a better solution, usually using branch swapping (Swofford & Olsen, 1990). The important point to note is that heuristic searches do not necessarily find the optimal solution, nor do they necessarily get particularly close to it. It is therefore important to accurately describe any heuristic procedures used in an analysis, so that the likely success of the method can be assessed.

There are many methods for calculating the distances for those methods that require them. Distances specify a relationship between pairs of taxa, with taxa that have identical states for all of the characters having zero distance between them, and taxa that have no character states in common having an arbitrarily large distance. Some data that are useful for cladistic purposes, such as those from immunology, nucleic acid hybridization and breeding experiments, are automatically expressed as pair-wise distances (West & Faith, 1980); but most character data need to be mathematically transformed into distances. The basic assumption when analysing distances is that the rate of nucleotide substitution varies from site to site according to a gamma frequency distribution (Jin & Nei, 1990). Part of the problem here is that the rate of nucleotide substitution varies within evolutionary lineages (Li et al., 1987) as well as among sites within a sequence (Olsen, 1987; Shoemaker & Pitch, 1989). Consequently, the nucleotide evolution models may be better if the variance of the estimates is reduced (Pollock & Goldstein, 1995), and if they assume that the rate of nucleotide substitution varies from site to site according to a gamma frequency distribution (Jin & Nei, 1990).

For DNA/RNA sequence data, the distance between taxa can be estimated by the total number of nucleotide differences between them (Gojobori, Moriyama & Kimura, 1990). However, rather than just count the number of observed differences in nucleotides between 2 taxa, it is usual to recognize that there are likely to have been many superimposed substitution/indel events through evolutionary time (e.g., multiple substitutions at any one nucleotide position), and that these superimposed events cannot be observed in the contemporary taxa. Many methods have thus been proposed to try to correctly estimate sequence divergence from observed nucleotide differences (see Gojobori et al., 1990; Pollock & Goldstein, 1995; Rzhetsky & Nei, 1995), based on models of nucleotide evolution that vary from very simple to quite sophisticated. These methods work by estimating the expected amount of nucleotide divergence between the sequences, assuming some specified probability that nucleotide changes will occur during an evolutionary time period (such as that the number of changes follows a poisson frequency distribution), usually by differentially weighting the various transitions and transversions that are possible among the 4 nucleotide character-states (A, C, G, T/U) to reflect their relative probability of occurrence. The differences observed between the taxa are thus "corrected" for unobserved multiple substitution/indel events, and the distances derived from these corrected data should provide a better estimate of the evolutionary divergence among the taxa.

The studies to date indicate that the simplest of the various models of nucleotide evolution that is likely to apply to real data sets is the 2-parameter model of Kimura (1980), but that it may not be applicable when the sequences are distantly related and/or the total number of nucleotides examined is large (Rzhetsky & Nei, 1995). Part of the problem here is that the rate of nucleotide substitution varies within evolutionary lineages (Li et al., 1987) as well as among sites within a sequence (Olsen, 1987; Shoemaker & Pitch, 1989). Consequently, the nucleotide evolution models may be better if the variance of the estimates is reduced (Pollock & Goldstein, 1995), and if they assume that the rate of nucleotide substitution varies from site to site according to a gamma frequency distribution (Jin & Nei, 1990).

Many other methods have also been proposed specifically for calculating distances from protein-coding DNA sequences, based on comparing codons rather than nucleotides. They recognize the distinction between non-synonymous and synonymous nucleotide substitutions (i.e. those that do and do not result in changes in the amino acid coded for), and include both weighted and unweighted pathways methods (Gojobori et al., 1990). Distances may also
be calculated directly from amino acid sequences (Ota & Nei, 1994). Furthermore, a number of methods have been proposed specifically for converting frequency data (such as allelic or genotypic frequencies) to distances, based on different genetic models (Rogers, 1991), and also for calculating distances from restriction endonuclease data (Nei & Li, 1979).

Tree-building methods. Here, I provide only a brief introduction to the procedures employed by the various tree-building methods. More details of most of these methods, along with worked examples, are provided by Swofford & Olsen (1990), and Charleston, Hendy & Penny (1994) provide a brief mathematical exposition.

The Unweighted pair-group method using arithmetic average (UPGMA) is an algorithmic method that analyzes distance data (Michener & Sokal, 1957; Sokal & Michener, 1958). It groups the taxa based on increasing distances between the taxa, starting by grouping the 2 taxa with the smallest distance, and then progressively adding more-distant taxa to the group or to new groups. It assumes that the data are ultrametric; that is, a tree can be constructed so that the observed distance between any 2 taxa (i.e. in the original data matrix) is equal to the sum of the branch lengths connecting them (i.e. from the original distance matrix) and the tree is rooted so that all of the taxa are equidistant from the root.

Neighbor-joining (Saitou & Nei, 1987; Studier & Keppler, 1988) is another algorithmic method that analyzes distance data. It is similar to UPGMA in that it starts by grouping the 2 taxa with the smallest distance, and then progressively adds more-distant taxa to the group or to new groups. However, it differs in modifying the distance matrix at each step, so that each pair of groups is adjusted on the basis of their average divergence from all other groups. The method thus relaxes the ultrametric assumption, and assumes only that the data are additive, that is, a tree can be constructed so that the observed distance between any 2 taxa (in the original distance matrix) is equal to the sum of the branch lengths connecting them on the tree (patristic distance or pathlength).

The Neighborliness procedure (Sattath & Tversky, 1977; Fitch, 1981) is the final algorithmic method discussed that analyzes distance data. It is based on evaluating the 4-point metric (Dubouloz, 1974), a relaxed definition of additivity for a set of 4 taxa. It specifies the conditions for clustering 4 taxa into 2 pairs, based on the relative ranking of the patristic distances between all possible pairs of the taxa. Neighborliness is thus a quartet method, evaluating the 4-point metric for all possible subsets of 4 taxa (quartets) to decide how the 4 taxa should be clustered, and then building clusters of the taxa that are consistent with the largest fraction of the quartets (i.e. the tree is the one for which the 4-point metric is satisfied the largest number of times).

Other algorithmic methods that analyze distance data have been described by Farris (1977), Chakraborty (1977), Klotz & Blanken (1981), Li (1981), Tateno, Nei & Tajima (1982), Hasegawa, Kishino & Yano (1985), and Charleston et al. (1994); but none of these has been widely used.

The Minimum-evolution procedure (Edwards & Cavalli-Sforza, 1964) analyzes distance data using an optimality criterion. The tree chosen is the one that minimizes the sum of the lengths of all of the branches on the tree, these branch lengths being estimated using linear algebra on the observed distances (i.e. from the original distance matrix) between 3 taxa inter-connected by a common node. Rzhetsky & Nei (1993) provide an exact method for optimizing each tree; and Rzhetsky & Nei (1992) have suggested an heuristic method for finding the optimal tree by pointing out that the neighbor-joining algorithm is guaranteed to find the Minimum-evolution tree for 4 taxa.

The Distance-Wagner method (Farris, 1972) also analyzes distance data using an optimality criterion. The optimality criterion is the same as for the Minimum-evolution procedure; the desired tree is the one that minimizes the sum of all of the branch lengths on the tree. However, in this case all of the patristic distances (along the tree branches) connecting any 2 taxa are constrained to be greater than the observed distances between these taxa, whereas the Minimum evolution procedure assumes that some of the patristic distances will be greater and some smaller than the observed distances. Exact methods for optimizing each tree have been described (Beyer et al., 1974; Waterman et al., 1977), and heuristic procedures for finding the optimal tree are usually used (Farris, 1972; Swofford, 1981; Faith, 1985).

The Weighted least-squares (Fitch & Margoliash, 1967) and the Unweighted least-squares (Cavalli-Sforza & Edwards, 1967) are related methods that analyze distance data using an optimality criterion. They both choose the tree that minimizes the error (i.e. the measured discrepancy) between the observed distances between the taxa (i.e. the original distance matrix) and the patristic distances connecting the taxa on the tree (pathlength). So, for each possible tree they compare the sum of all of the squared differences between the distances in
the original matrix and those on the tree, and then choose the tree that minimizes this sum. The 2 methods differ in the formula they actually use to measure the discrepancy between the observed and patristic distances, the unweighted method assuming that all of the distance estimates have the same absolute error and the weighted method assuming that they have the same percentage error. Exact methods for optimizing each tree have been described (Kidd & Sgaramella-Zonta, 1971; Beyer et al., 1974; Waterman et al., 1977), and heuristic procedures for finding the optimal tree are usually used.

The Maximum-parsimony method (Farris, 1970; Fitch, 1971) is the first method to be discussed that analyses the original data matrix using an optimality criterion. It is probably the most popular method among cladists, and indeed many cladists restrict the definition of cladistics to include this method alone. The optimal tree is the one that has the minimum number of inferred character state changes (i.e. shortest tree length or fewest evolutionary steps). Exact methods for optimizing each tree have been described (Fitch, 1971; Swofford & Maddison, 1987). Exact methods for finding the optimal tree include an exhaustive search (i.e. testing every possible tree topology) and branch-and-bound methods, but heuristic methods must be used for analyses with more than about 20 taxa (Swofford & Olsen, 1990).

Special methods have been developed for frequency data, such as allelic or genotypic frequencies (Swofford & Berlocher, 1987).

The Weighted-parsimony method (Farris, 1969; Sankoff, 1975) is a generalization of the Maximum-parsimony method that allows the characters and character states to be weighted so that they contribute unequally to the calculation of optimality (note that character weighting and character-state weighting are separate ideas; Williams & Fitch, 1989). The optimal tree is the one that has the minimum sum of the weights. The rationale for this is that the unequal weights reflect unequal probabilities of change in the characters (Felsenstein, 1981b), the weighting usually being applied so that evolutionarily conservative characters or states can contribute more to the reconstruction of the phylogeny (Olsen, 1987). The magnitude of the weights to be applied can be decided either a priori (i.e. decided before the analysis) or a posteriori (i.e. calculated after an initial cladogram has been constructed) (Neff, 1986).

Several general a priori weighting schemes have been proposed, such as Dollo-parsimony (Le Quesne, 1974), where no parallelisms or convergences are permitted, and Camin–Sokal-parsimony (Camin & Sokal, 1965), where no reversals are allowed, as well as compatibility weighting (Penny & Hendy, 1986; Sharkey, 1989). However, many a priori weighting schemes have also been developed specifically for molecular sequences, where the genetic mechanisms responsible for nucleotide substitutions make it seem unreasonable to assume equal weights for all changes. Differential weights can be applied to (see Hillis et al., 1993): (1) different sequence positions (character weighting), thus emphasizing functional inequalities along the sequences (e.g., different codon positions in translated genes, stems vs loops in structural RNAs, hypervariable positions in mtDNA); or (2) the same sequence positions (character-state weighting), thus emphasizing mutational biases (e.g., transitions vs transversions, relative substitution frequencies, base composition, synonymous vs non-synonymous changes). A posteriori weighting schemes include successive approximations (Farris, 1969), quadratic weighting (Fitch & Yasunobu, 1975), transformation series (Micko & Fitch, 1982), dynamic weighting (Williams & Fitch, 1989) and homoplasy weighting (Goloboff, 1993). Exact methods for optimizing each tree using weights have been described (Sankoff & Cedergren, 1983).

The Compatibility method (Le Quesne, 1969, 1982) also analyses the original data matrix using an optimality criterion. The optimal tree is the one that has the largest set (clique) of mutually compatible characters (i.e. characters for which there is no homoplasy on the tree). The method proceeds by testing the pair-wise compatibility of all of the characters, and then examining all possible cliques of characters. The method thus relies on the theorem that a group of characters which are pair-wise compatible are jointly compatible. This is always true for both polarized and unpolarized 2-state characters (Estabrook & McMorris, 1980), but for unpolarized multi-state characters it is possible for a collection of pair-wise compatible characters not to have any phylogeny for which they are all compatible (Fitch, 1975). Branch-and-bound methods are available for finding the maximal cliques.

The Maximum-likelihood procedure (Felsenstein, 1973a,b) is another method that analyses the original data matrix using an optimality criterion. So far, practical likelihood methods have only been developed for analysing nucleotide sequence data (Felsenstein, 1981a), protein sequence data (Kishino, Miyata & Hasegawa, 1990), and restriction site data (Felsenstein, 1992), with restricted maximum-likelihood methods also available for continuous characters (Felsenstein, 1981c). The optimal tree is the one that maximizes the statistical likelihood that the specified evolutionary model produced the observed data. For example, for a given model of nucleotide evolution (the most commonly-used one is
the generalized Kimura 2-parameter model described by Kishino & Hasegawa, 1989), formulae are derived that describe the probability that an initial nucleotide will be transformed into a specified nucleotide during an evolutionary time-period. The likelihood for each nucleotide position is then equal to the prior probability of finding the initial nucleotide at that position multiplied by the probability of transformation. The likelihood of the divergence of 2 sequences during the time-period is then the product of the likelihoods at each position, and the overall likelihood for a tree is the product of the likelihoods along the branches. No exact methods for optimizing each tree have been described, and iterative (“hill-climbing”) methods are used (Fukami & Tateno, 1989). Exact methods for finding the optimal tree involve an exhaustive search (since suitable branch-and-bound methods are unknown), and so heuristic methods (Felsenstein, 1989a) are usually used for analyses with more than 5 taxa.

The Invariants (or Evolutionary parsimony) method (Lake, 1987; Cavender & Felsenstein, 1987) also analyses the original data matrix using an optimality criterion, although it was developed specifically for analysing nucleotide sequence data. It is based on evaluating the number of operator invariants (linear for the method of Lake; quadratic for the method of Cavender & Felsenstein) for 4 taxa, which specifies the conditions for clustering the 4 taxa into 2 pairs, based on the relative number of transversions on the internal branch of the unrooted tree to those on the terminal branches. The use of Invariants is thus a quartet method, evaluating the invariants for all possible subsets of 4 taxa (quartets) to decide how the 4 taxa should be clustered, and then building clusters of the taxa that are consistent with the largest fraction of the quartets (Lake, 1988). Exact methods are used for the evaluation of the invariants.

Spectral analysis (Penny, Hendy & Henderson, 1987; Hendy & Penny, 1993) is the final procedure that analyses the original data matrix using an optimality criterion. It has been developed only for nucleotide data, originally with 2-state characters (Hendy, 1989) and more recently with 4 states (Hendy, Steel & Penny, 1994). This procedure first uses the hadamard transformation (Hendy & Charleston, 1993) together with a model of sequence evolution (such as the 2-character-state model of Cavender, 1978a,b) to adjust (or correct) the observed data for unobserved character-state changes (e.g., multiple substitutions at a sequence position), thus producing a matrix of expected data (i.e. expected number of character-state changes under the evolutionary model adopted). These corrected data are then used to estimate the degree of support for every possible subset of taxa, by summing the number of the expected character-state changes that are compatible with each subset. The Closest-tree method (Hendy, 1991; Steel et al., 1992), a compatibility-type method based on minimizing the sum-of-squares, is then used to select the optimal tree as the one with the largest set of mutually compatible subsets of taxa. Exact methods for finding the optimal tree include branch-and-bound methods (Hendy & Penny, 1989), with a limit of about 20 taxa.

Characteristics. Clearly, it is impossible to compare the many tree-building methods based on their relative ability to detect the phylogenetic pattern underlying the apparent incongruences in real data sets, because the real phylogenetic pattern is almost never known. Instead, many other criteria have been proposed to assess the relative usefulness of the various tree-building methods, although many of these criteria have not yet been effectively evaluated for most of the methods. Those criteria about which we know the most are outlined here — no single method meets all of the criteria (Penny et al., 1990, 1992).

The Computational Efficiency of a method refers to the relative speed with which a solution is produced (Penny et al., 1992). All of the algorithmic procedures are relatively efficient, as the computation time does not increase dramatically as the number of taxa is increased. On the other hand, none of the optimality methods is efficient, because the number of trees from which the optimal one needs to be chosen (i.e. at the second step in the double-optimization) increases exponentially as the number of taxa increases (Felsenstein, 1978a). Finding the optimal tree using exact methods is an example of the large class of NP-complete mathematical problems (Graham & Foulds, 1982), for which no efficient procedure is known (or is expected ever to be known). So, for large data sets (both in terms of number of taxa and number of characters) heuristic methods must be used for all of the optimality methods if a solution is to be produced within a reasonable time. This means that finding the optimal solution cannot be guaranteed for the analysis of large data sets — just how large “large” is depends on the computing power (both software and hardware) available for the analysis and the patience of the cladist. This situation is compounded for data sets with a relatively low phylogenetic signal (see below), which are likely to have multiple optimal trees that are structurally quite different (Maddison, 1991) — these “islands” of trees may not easily be found by heuristic techniques (Page, 1993).
The Maximum-likelihood method is even more inefficient than all of the other methods that use optimality criteria. For Maximum-likelihood, the first step in the double-optimization (i.e. optimizing each tree topology) has no simple analytical solution (Fu & Takano, 1989), unlike all of the other methods, and so computationally-intensive iterative procedures must be used. This problem is exacerbated by the fact that there may be more than 1 way of optimizing a particular tree topology (Steel, 1994). These problems are then compounded by the lack of a branch-and-bound algorithm for the second step in the double-optimization; and so there is a severe limit to the size of the data set that it is practical to analyse by the Maximum-likelihood method.

The Power of a method refers to the response of the method to increasing amounts of data (Olsen, 1988) — with an increasing number of characters in the original data matrix we expect a method to converge to a single tree (i.e. a limit is reached beyond which further increases in the data do not change the tree produced by the method) (Penny & Hendy, 1986). A powerful method is one in which the convergence occurs with a relatively small number of characters (Hillis, Huelsenbeck & Swofford, 1994b). Power is also related to the amount of information in the original data matrix that is actually used to construct the tree (Penny et al., 1992, 1993) — all of the tree-building methods omit some of the character information, and so the power of these methods may be reduced. However, it is important to note that not all of the character information is necessary for the reconstruction of an evolutionary tree, and that therefore it is not loss per se but loss of relevant information that is important (Shoemaker & Fitch, 1989; Penny et al., 1993).

Compared with the character-based methods, all of the distance methods involve a very large loss of information in converting the original data into distances. This is because there will be multiple character data sets that produce exactly the same distance matrix (Felsenstein, 1982; Penny, 1982), and this problem becomes worse for increasing numbers of taxa and characters (Steel, Hendy & Penny, 1988). Moreover, for sequence data it is not clear how to calculate distances when the aligned sequences contain terminal length variations, indels or ambiguous nucleotides (since the models of nucleotide evolution are based on estimating rates of transitions and transversions only) — including or excluding these features in the calculations can produce different distance estimates (Swofford & Olsen, 1990; Beanland & Howe, 1992).

The character-based methods also do not use all of the data (Penny et al., 1990). For example, the quartet methods ignore that part of the information that comes from the incompatibilities among the quartets, while Compatibility analysis assumes that incompatible characters have no phylogenetic information at all (Hill, 1975). Furthermore, the Parsimony and Compatibility methods do not use characters that have constant states across all of the taxa, nor characters where only 1 of the states occurs in more than 1 taxon (autapomorphies); and these may form a large proportion of sequence data (for example, I looked at an arbitrary selection of 20 molecular studies published in 1994, and found that on average 84% of the nucleotide positions were uninformative for Parsimony/Compatibility analysis using these 2 criteria). The Maximum-likelihood, Invariants and Spectral-analysis methods ignore characters with missing character-state data for any of the taxa, including indels for sequence data, which may also constitute a sizable proportion of the original data matrix. The Invariants method is also based entirely on transversion substitutions, and there may not be enough in a data set to effectively infer the branching pattern, or the frequency of transitions may confound the pattern (Jin & Nei, 1990; West & Faith, 1990); it thus appears to have very low power (Hillis et al., 1994a,b).

The Consistency of a tree-building method refers to its ability to converge to the correct tree (Felsenstein, 1988). That is, with an increasing number of characters in the data matrix a consistent method converges to the single correct tree, and a method is inconsistent if there are any possible situations where it converges to the wrong tree (i.e. there are systematic or non-random errors). Consistency is related to the assumptions (e.g., the evolutionary model) made by the tree-building method — if the assumptions are not met for a particular data set then consistency cannot be expected.

There are fundamental assumptions underlying all of the tree-building methods, such as that evolution proceeds as a continuous-time markov process (Beanland & Howe, 1992), whether they are explicit or not for a particular type of data (e.g Rodriguez et al., 1990). The most notable assumption is that character-state changes are independent and identically distributed (Cavender, 1978a; Olsen, 1987; Shoemaker & Fitch, 1989; Sanderson, 1995); which means that no character-state change affects any other character-state change, and that all character-state changes are equally likely. If evolution has occurred then this assumption is generally not true, because of the common history of the taxa during at least some evolutionary period. The distance methods will be particularly sensitive to this problem.
(Felsenstein, 1988), although all of the methods suffer from it (Felsenstein, 1982).

The distance methods will also be inconsistent if the data are not additive (Felsenstein, 1988), as there is no exact solution for estimating the tree under these circumstances (Swofford & Olsen, 1990). Additivity of taxonomic data is unlikely because the distances are only estimates of evolutionary divergence, and are therefore expected to vary randomly around the true divergence value (i.e. the data will be additive only if there is no homoplasy). One of the consequences of this is that trees with negative branch lengths are theoretically possible when using the Neighbor-joining, Minimum-evolution and Least-squares methods (Swofford & Olsen, 1990). UPGMA adds the further constraint that the data are ultrametric, that is, that evolutionary divergence among the taxa occurs at a uniform rate through time (such as occurs with a molecular clock). This is also unrealistic.

All of the methods will also be inconsistent if the amount of homoplasy is large, because homoplasy makes unrelated taxa appear to be more similar (Olsen, 1987; Swofford & Olsen, 1990). This will be a particular problem if the frequency of character-states varies between taxa (Penny et al., 1990); for example, for sequence data similar GC-content often unites unwanted branches on a tree (Siddow & Wilson, 1990, 1991; Lockhart et al., 1992; Marshall, 1992; Hasegawa & Hashimoto, 1993), and all tree-building methods are sensitive to this (Lockhart et al., 1992, 1994).

Methods may also be inconsistent when there are juxtaposed long and short branches on the tree (e.g., unequal rates of phylesis or speciation), as these branches will tend to join together (often referred to as “long branches attracting”). This has long been recognized as a particular problem for the Parsimony and Compatibility methods (Cavender, 1978a; Felsenstein, 1978b; Hendy & Penny, 1989; Zharkikh & Li, 1993; Takezaki & Nei, 1994), but we now recognize that it also applies to all of the distance methods as well as to the Invariants and Closest-tree methods (Jin & Nei, 1990; Penny et al., 1990; de Bry, 1992; Huelsenbeck & Hillis, 1993; Steel, Hendy & Penny, 1993; Zharkikh & Li, 1993; Hillis et al., 1994b). The Maximum-likelihood method appears to suffer least from this problem (Hillis et al., 1994b).

The Robustness of a tree-building method refers to how much the assumptions of the method can be violated before the method becomes inconsistent (Penny et al., 1990). This may be the most important criterion, because the idealized assumptions of the methods are likely to be violated by real data sets (Huelsenbeck, 1995a,b). Many of the consistency problems listed above may thus not necessarily preclude use of the various methods if they are robust.

The distance methods, for example, may be quite robust to the problem of non-additivity if the distance estimates are corrected for multiple character-state changes and the problems of negative branch-lengths are addressed (Felsenstein, 1988); and even UPGMA does not require perfectly ultrametric data (Colless, 1970). The consistency problems associated with homoplasy can also potentially be solved (i.e. the methods can be made robust) by any method that provides a framework for making corrections for multiple character-state changes (Penny et al., 1993), such as the use of transformations and corrected distances (e.g., for unequal nucleotide composition, use of the LogDet transformation, Lockhart et al., 1994; compositional statistics, Siddow & Wilson, 1990, 1991; or character weighting, Marshall, 1992). Consistency can often be restored (i.e. the methods made robust) when there are juxtaposed long and short branches by adding taxa to the analysis that join the tree on the long branches (Hendy & Penny, 1989; Penny et al., 1990; Lento et al., 1995), or by using character weighting schemes that shorten the branches (Swofford & Olsen, 1990; Hillis et al., 1993). However, the preferred method is to make corrections for multiple character-state changes using transformations (Jin & Nei, 1990; de Bry, 1992; Steel et al., 1993; Huelsenbeck, 1995a). Note that the attempts to improve the consistency of the tree-building methods by using corrections and transformations apply only to molecular data, because it is possible to use a consistent model of character-state changes that will apply to all of the characters in any 1 data set. It is not obvious how to appropriately correct or transform other types of data, since each character will have its own model of character-state change.

A popular approach to studying the robustness of the various methods is by using simulated data, where the original tree and the characteristics of the data are known, and the behavior of the tree-building methods under these conditions is evaluated (e.g., Tateno et al., 1982; Li et al., 1987; Sourdis & Krimbas, 1987; Kim & Burgman, 1988, Saitou & Imanishi, 1989; Jin & Nei, 1990; Rohlf et al., 1990; Nei, 1991; Siddow & Wilson, 1991; Hasegawa & Fujiwara, 1993; Huelsenbeck & Hillis, 1993; Charleston et al., 1994; Hillis et al., 1994a,b; Kuhner & Felsenstein, 1994; Tateno, Takezaki & Nei, 1994; Gaut & Lewis, 1995; Huelsenbeck, 1995a,b; Tillier & Collins, 1995; Yang, 1995). There is clearly a large conceptual gap between simulated data and real data (Miyamoto & Cracraft, 1991; Hillis et al., 1994a), and most of the simulations apply only to the
branching order of a limited number of taxa under specific evolutionary models. The jury is still out on the results (cf. Felsenstein, 1988), but these studies show that many of the tree-building methods are actually quite effective (i.e. robust) over a wide range of evolutionary conditions. The most robust of the methods appears to be the Maximum-likelihood procedure, with UPGMA and the Invariants method being the least robust. The robustness of the other methods seems to depend very much on the ability to correctly utilize transformations for multiple character-state changes or to appropriately apply character or character-state weighting; clearly this requires a priori knowledge about the evolutionary process in the taxa being studied that is unlikely to be available. It is unfortunate that Spectral analysis has not been incorporated into most of the simulation studies to date, in order to test whether its theoretical advantages are met in practice.

Finally, there are methodological and philosophical considerations that are often used to compare the techniques (Miyamoto & Cracraft, 1991). In particular, the algorithmic procedures are often criticized because they provide no method for ranking alternative tree topologies (i.e. there is no criterion to compare trees and decide how “good” any of them are as estimates of phylogeny) (Swofford & Olsen, 1990). This is a serious limitation if we treat a cladogram as being an attempt to estimate the true phylogeny, because we are using incomplete information and there is thus a degree of uncertainty about the correctness of our estimate (Felsenstein, 1988). It is thus always a good idea to assess how much better the optimal tree is than the next-most-optimal tree (or set of trees) (Swofford, 1991), especially as these trees may not be independent of each other (Penny et al., 1995).

After tree construction. We can never know whether our cladogram actually represents the true phylogenetic history of the taxa concerned. Therefore, there are 3 further issues that need to be dealt with once a phylogenetic tree has been constructed using 1 or more of the methods outlined above: assessing the magnitude of the phylogenetic signal in the cladogram; assessing the robustness of the monophyletic groups represented on the cladogram; and comparing the structure of multiple cladograms from the same set of taxa. Although these issues are extremely important for cladistic analysis, they are only briefly introduced here because a full coverage would require a separate (although shorter) review.

Even randomly-generated data will lead to the construction of a cladogram if any of the tree-building techniques is applied to them. It is therefore important to quantitatively assess the magnitude of the phylogenetic signal (i.e. that part of the character variation that is potentially informative about the evolutionary history) in a set of data — is the support for a particular tree any better than would be expected from a random data set? Quantifying this information also allows cladograms to be compared effectively with each other — a cladogram with a stronger phylogenetic signal is to be preferred. Four classes of techniques have been developed: optimality measurements (for those methods that use an optimality criterion), such as consistency and homoplasy indices (Archie, 1989; Farris, 1989, 1991) for the Parsimony methods, likelihood values for the Maximum-likelihood methods, and sums-of-squares for the Least-squares methods; skewness indices (Huelsenbeck, 1991; Hillis & Huelsenbeck, 1992); analytical statistical tests (Felsenstein, 1988; Li & Gouy, 1991; Li & Zharkikh, 1995); and randomization (or permutation) statistical tests (Goloboff, 1991; Källersjö et al., 1992; Alroy, 1994; Faith & Ballard, 1994).

A cladogram is interpreted in terms of the monophyletic groups of taxa that it hypothesizes. It is therefore important to quantitatively assess the robustness of all of the groups indicated by the branching sequence of a cladogram (i.e. the degree of support for each branch in the tree) — is the support for a particular group any better than would be expected from a random data set? Many classes of techniques have been developed, including: analytical procedures, such as confidence limits, branch-length variances and likelihood-ratio tests (Li & Gouy, 1991; Li & Zharkikh, 1995); resampling procedures, such as the bootstrap (Li & Zharkikh, 1994, 1995; Sanderson, 1995), the jackknife (Felsenstein, 1988) and topology-dependent permutation (Faith, 1991); and non-statistical procedures, such as the decay index (Bremer, 1994), clade stability (Davis, 1993), and spectral signals (Hendy & Penny, 1993).

It is unlikely that different tree-building methods will always produce identical trees for the same data set, given their different underlying assumptions, and it is also possible for multiple optimal trees to be produced by a single tree-building method. Furthermore, we may wish to compare cladograms produced using different sets of characters (since congruence among data sets may constitute the strongest achievable evidence for the true phylogeny; Penny & Hendy, 1986), and these may differ as a result of sampling error, different stochastic processes, or different branching histories (de Queiroz et al., 1995). It is therefore usually necessary to quantitatively compare the structure of multiple trees from the same set of taxa. For example, which monophyletic groups
A parasitological example. To explore the practical consequences for phylogeny reconstruction of the wide range of available tree-building methods, it is worthwhile exploring a concrete example. This parasitological example illustrates and emphasizes many of the points made in earlier sections, as well as providing a brief step-by-step guide to a cladistic analysis.

Methods. First, we need to consider the background to the problem that we intend to address with the cladistic analysis. The members of the genus Sarcocystis are parasitic protozoans with heteroxenous life cycles, in which both hosts are mammals. The sexual stage is in the definitive host (often a predator species) and the asexual stage is in the intermediate host (often a prey species), with the sporocysts being passed in the faeces of the definitive host and ingested by the intermediate host. Studies of some of the species are important pathogens of livestock and humans.

The current classification of the phylum Apicomplexa (Sporozoa) is based largely on ultrastructural analysis, on details of the life-cycles, and on the location of developmental stages. The internal arrangement of the phylum is currently rather unstable, at least partly because of a lack of knowledge of the life-cycles for many taxa. Cox (1994) treats the coccidial parasites as a class (the Coccidea), with Sarcocystis in the order Eimeriida. The family Sarcocystidae is often subdivided into 2 subfamilies, the Sarcocystinae (3 genera, 100 species) and the Toxoplasmatinae (2-3 genera, 15 species) (Levine, 1985).

However, the phylogenetic relationships within the Sarcocystidae remain ambiguous, at least partly because of difficulties in assessing the homology of phenotypic characters. It is possible that molecular data might alleviate these problems, although there has been little consistency among molecular studies to date. While the family as a whole is generally considered to be monophyletic, several analyses of small-subunit ribosomal RNA (ssrRNA) sequences have questioned the monophyly of Sarcocystis itself (Barta, Jenkins & Danforth, 1991; Tenter, Baverstock & Johnson, 1992; Ellis et al., 1994). In particular, these studies suggest a strong correlation between the parasite phylogeny and that of their definitive hosts, with species of the Sarcocystidae forming 2 clades based on the use of either canids or felids as their definitive hosts. Thus Sarcocystis (felid or canid hosts) is only monophyletic if both Toxoplasma (felid host) and Neospora (unknown host) are included within it.

Second, we need to consider the data that are available to tackle this problem (i.e. the ingroup). There are currently complete ssrRNA sequences available for 8 species of the Sarcocystidae (Sarcocystis arieticanis, S. fusiformis, S. gigantea, S. muris, S. neurona, S. tenella, Neospora caninum, Toxoplasma gondii), which can be used to test the monophyly of Sarcocystis. Several published sequences are available for some of the species (9 for T. gondii, 2 for N. caninum), and for the purposes of this example consensus sequences were derived for these species (after alignment; see below) using the MacClade 3.04 computer program (Maddison & Maddison, 1992). The standard IUPAC ambiguity codes were used for those few nucleotide positions with more than 1 possible character-state in the sequence (1 for S. muris, 1 for N. caninum, 26 for T. gondii). Thus, we have a representative sample of the family (e.g., species with both felid and canid hosts), and we have taken inter-individual variation into account as far as possible.

Third, we need to consider how the cladogram will be rooted (i.e. the outgroup), since monophyly can be assessed only on a rooted tree. The members of the Eimeriida that are most closely-related to the Sarcocystidae (see Levine, 1985) are the members of Eimeria (family Eimeriidae), for which 7 species have published ssrRNA sequences (E. acervulina, E. brunetti, E. maxima, E. mitis, E. necatrix, E. praecox, E. tenella). For the outgroup, a single consensus sequence was derived from the sequences for these 7 Eimeria species (after alignment) using the MacClade program, with the standard IUPAC ambiguity codes being used for those 68 nucleotide positions with more than 1 character-state in the consensus sequence. These ambiguous character-states can be dealt with effectively by some analysis techniques, while others treat them as missing data. For this data set this different treatment has very little effect, because in almost all cases the Sarcocystis sequences are invariant for these characters (i.e. the variability is only within the Eimeria phylogeny itself and it thus does not influence the reconstruction of the Sarcocystis phylogeny).

Fourth, we need to consider character homologies (i.e. sequence alignment). There are 9 sequences in...
the analysis, 8 for the ingroup and 1 for the out-group. The 18SrRNA sequences for all of these species were aligned using the secondary structure model of Van de Peer et al. (1994) — all of the sequences were obtained from this computer database. For the cladistic analysis, the single-stranded regions were omitted, as Ellis & Morrison (1995) have indicated that it is the double-stranded region that contains most of the phylogenetic information for these taxa. There are thus 694 nucleotide positions in the analysis. This procedure effectively deals with both the problems associated with positions of equivocal alignment (since there are very few indels in the double-stranded region), as well as those associated with relative character weighting based on structural/functional inequalities along the sequences.

Fifth, we need to consider the tree-building methods. Four tree-building methods were used (covering the full range of possibilities), in order to test the sensitivity of the analysis to this source of variability. Two of these were distance-based methods, 1 of them algorithmic (Neighbor-joining) and the other with an optimality criterion (Unweighted least-squares). The other 2 methods were character-based with optimality criteria, one developed specifically for nucleotide sequences (Maximum-likelihood) and one not (Weighted-parsimony). The Phylip 3.57 computer program (Felsenstein, 1995) was used for the first 3 methods, employing the heuristic search with global branch-swapping for those methods with an optimality criterion. The PAUP 3.1.1 computer program (Swofford, 1993) was used for the final method, employing the branch-and-bound search. Note that in order to test the monophyly of Sarcocystis, the only part of the cladogram that needs to be interpreted is the branching order (not the branch lengths), since we will be independently assessing the degree of support for monophyly (see below).

Sixth, we need to consider the evolutionary model or models that we wish to employ. Two models of character-state change were used, in order to test the sensitivity of the analysis to this source of variability. The first model assumed that all nucleotide changes were equally probable at each sequence position, while the second model assumed that transitions were twice as likely as transversions (for the trees shown in Fig. 5A the transition : transversion ratio is about 1.5, as determined by the MacClade program). For the maximum-likelihood method, the transition : transversion ratio was set as part of the specified model. For the distance-based methods, the Kimura distance measure was used to estimate sequence divergence, and the transition : transversion ratio was set to either 1 or 2. For the Parsimony method, a step-matrix was used, with transitions weighted as 1 and transversions weighted as either 1 or 2. The nucleotide frequencies are approximately equal for all of the sequences, and so there is probably no need to correct for GC-bias (the GC content varies from 50 to 60% across the sequences).

Next, we need to consider the magnitude of the phylogenetic signal in the data (i.e. the robustness of the cladograms). For those methods with an optimality criterion, permutation tests were used. Two-hundred permutations of the data were produced for each model and each tree-building method, and the percentage of these replicates that produced trees that were at least as optimal as the observed tree was calculated. The Phylip program implements this option for its tree-building methods. For the Parsimony analyses, some of the permutations were performed using the RandomCladistics 3.0 computer program (Siddall, 1994); for the remainder, the Phylip program was used to create the permuted replicates and these were then individually fed into the PAUP program.

Finally, we need to decide how we are going to assess the magnitude of the support that the data have for the monophyly of Sarcocystis. Bootstrap resampling was used, which is implemented in both the Phylip and PAUP programs. Two-hundred bootstrap replicates were produced for each model and each tree-building method, and the percentage of the trees derived from these replicates that indicated monophyly of Sarcocystis was calculated.

Results. Three cladograms resulted from the 8 analyses (Fig. 5), with each of the optimality methods finding a single most-optimal tree for each analysis. The Neighbor-joining and Maximum-likelihood analyses produced the same tree irrespective of the model used (Table 3), this tree supporting monophyly of Sarcocystis.

However, the Weighted-parsimony method produced different trees for the 2 models (Table 3), both supporting monophyly of Sarcocystis, but differing in the relative placement of S. muris and S. neurona. However, for the model in which transitions and transversions were equally probable, the tree in Fig. 5A is only one step longer (i.e. one more character-state change) than the tree in Fig. 5B, and so there is actually very little conflict between the 2 methods. This result emphasizes the importance of considering sub-optimal trees in a cladistic analysis.

The Unweighted least-squares method produced the same tree irrespective of the model used (Table 3), but this tree does not support monophyly of Sarcocystis, as S. muris and S. neurona are removed from the clade. The contradiction of monophyly in
Fig. 5. The 3 phylogenetic trees produced from the cladistic analysis of Sarcocystis ssrRNA. Table 3 shows which analysis produced which of these trees. Trees (A) and (B) support monophyly of Sarcocystis, whereas tree (C) does not.

This cladogram is supported only by a branch with negative length (i.e. the branch from the ancestor of the ingroup to the ancestor of T. gondii+N. caninum+S. gigantea+S. fusiformis+S. tenella+S. arieticanis). This situation is not realistic, and so the tree cannot be considered seriously as a possible representation of the real evolutionary history of these taxa. If this tree-building method is constrained to use only positive branch lengths (which is an option in the Phylip program), then the cladogram produced does indeed support monophyly of Sarcocystis using both models (they produce the cladogram shown in Fig. 5A).

The permutation tests indicate that there is very strong phylogenetic structure in these ssrRNA data, as no permutations were found that produced trees that were anywhere near as optimal as those produced by the original data set (Table 3). The phylogeny of Sarcocystis can thus be usefully evaluated using this data set.

The bootstrap resampling indicates relatively strong support for the Sarcocystis clade (i.e. Sarcocystis is monophyletic in more than 85% of the replicates) for all of the methods in which it occurs (Table 3). This support is apparently stronger (i.e. the bootstrap percentage is higher) when transitions are considered to be more likely than transversions. Even the Unweighted least-squares method provides some support for monophyly, because the negative branch-length that contradicts the monophyly is actually quite short.

It is thus reasonable to conclude from this analysis that the ssrRNA sequence data do support monophyly of Sarcocystis, and thus also support the way in
Table 3—Results of the cladistic analysis of *Sarcocystis* ssrRNA

<table>
<thead>
<tr>
<th>Tree-building method</th>
<th>Optimality measure</th>
<th>Tree</th>
<th>Permutation percentage</th>
<th>Bootstrap percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbor-joining</td>
<td>( t_{\text{tv}} )</td>
<td>A</td>
<td>( c )</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{tv}} )</td>
<td>A</td>
<td>( c )</td>
<td>91.5</td>
</tr>
<tr>
<td>Unweighted least-squares</td>
<td>( t_{\text{tv}} )</td>
<td>0.00241</td>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{tv}} )</td>
<td>0.00335</td>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>Weighted-parsimony</td>
<td>( t_{\text{tv}} )</td>
<td>329e</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{tv}} )</td>
<td>473b</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>Maximum-likelihood</td>
<td>( t_{\text{tv}} )</td>
<td>(-2439.7)</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{tv}} )</td>
<td>(-2502.1)</td>
<td>A</td>
<td>0</td>
</tr>
</tbody>
</table>

\( t_{\text{tv}}, \) transitions and transversions equal; \( t_{\text{tv}} \), transitions twice as likely as transversions.

As shown in Fig. 5.

Percentage of the permutation replicates that produced trees at least as optimal as the observed tree.

Percentage of the bootstrap replicates that support monophyly of *Sarcocystis*.

No optimality criterion, and so no calculations possible.

Sum of squares.

Number of character-state changes.

Number of weighted character-state changes.

Log-likelihood.

which the Sarcocystidae is currently divided into 2 subfamilies (if the recently-described *Neospora* is included in the Toxoplasmatinae) (note that in a Popperian sense the data actually “fail to disprove” the hypothesis rather than “support” it). This conclusion is robust both to variations in the evolutionary model and to the tree-building method used. It is not, however, necessarily robust to the sequence alignment used (see Ellis & Morrison, 1995).

There does, however, seem to be some uncertainty with respect to the placement of *Smuris* and *S. neurona* (which is the source of non-monophyly detected by the previous studies of ssrRNA referred to above). This uncertainty is only likely to be resolved by including in the analysis other species that are closely-related to these 2 species (i.e. that will join the tree on the terminal branches leading to these 2 species). This is the obvious direction to take for future research into the question posed for the analysis presented here.

Some advice for phylogeny reconstruction. The following is a list of topics that need to be considered when carrying out a cladistic analysis. The topics are more-or-less in the order in which they will need to be considered in practice.

Include in the analysis all of the taxa for which there are data. Adding or deleting any taxon may affect the outcome of the analysis. This is one of the biggest sources of problems when analysing molecular data.

Include samples from more than 1 individual or population if possible. This is the norm when assessing morphological characters, because if intra-taxon variability exists then it will be confounded with inter-taxon variability in the analysis. For molecular data, a consensus sequence can be used for each taxon.

The ingroup must be monophyletic for the cladogram to accurately reflect phylogenetic history. If the analysis shows that the ingroup and outgroup taxa are inter-mixed in the cladogram, then it is not possible to root the cladogram in such a way that the ingroup is monophyletic. You should thus seriously consider your original rationale for assuming monophyly.

Long branches on a cladogram (indicating apparently extensive phylesis) should be avoided if this is possible. They can often be shortened by including taxa that join the tree on these branches (i.e. the branches actually show speciation rather than phylesis, but they are not doing so because the relevant taxa are not in the data set).

The outgroup should consist of more than 1 taxon closely-related to the ingroup. The more distantly-related is the outgroup then the more arbitrary is the apparent position of the root on the cladogram.
For molecular data, use entire sequences if they are available. There is no reason to expect part of a gene sequence (effectively a sample) to reflect accurately the total gene sequence, unless all of the nucleotide positions agree on the phylogenetic branching order. Homology assessments should be taken very seriously. For molecular data, this means that sequence alignment should be given careful consideration, using models of secondary structure if they are available. If computer programs are used, then vary the weighting parameters available in order to detect those parts of the sequence alignment that are not robust, and delete these regions from the analysis.

Assess the characteristics of the data before the cladogram is constructed (e.g., GC-content for nucleotide data), in order to choose rationally among the available tree-construction algorithms.

Use several cladogram-construction algorithms if this seems to be appropriate for the data. If heuristic tree-search methods are used then describe them precisely, so that the likely success of having found the optimal tree can be assessed. For nucleotide data, correct for multiple character-state changes if possible.

Assess the magnitude of the phylogenetic signal in the cladogram. Assess the robustness of the monophyletic groups represented on the cladogram. Quantitatively compare the cladograms if there is more than one (e.g., multiple optimal trees, trees from different cladogram-construction algorithms, or trees from different data sets).

Remember that the cladogram represents a series of hypotheses that are amenable to further test. It is not a representation of reality, nor is it the final word on the phylogeny of the taxa analysed.

Conclusion. Many biologists treat data analysis in a manner that is completely antithetical to their apparent attitude to the other, more biological aspects of their scientific pursuits (cf. Felsenstein, 1988). In particular, there is a widespread view among non-experts that there should be only one correct method for phylogenetic data analysis, and that it is about time that the experts told us what it is. This view is clearly rather naive, because it assumes that all phylogenetic data sets have identical characteristics. This is patently untrue, as it will be for all data sets (whether biological or not) that incorporate an historical component. Furthermore, currently no methods are simultaneously efficient, powerful, consistent and robust (Penny et al., 1990, 1992), and so there is no single ideal method.

The future development of cladistic techniques will presumably follow the well-documented path taken by statistics, in which a series of mathematical tests have been developed (and are still being actively developed) that make different assumptions about the nature of the data to be analysed, so that after the data have been evaluated an informed decision can be made about which test is the most appropriate for the particular task at hand. Thus, out of the current plethora of phylogenetic techniques should emerge a subset of mathematical methods that are known to be robust and consistent under certain circumstances, and it will be from amongst these methods that an objective choice (based on efficiency and power) will be made by the cladist.

The current vogue for simply presenting the results of analyses using several tree-building methods is thus both creditable and naive. It is creditable because it emphasizes our current uncertainty about cladistic analysis; and it is naive because there appears to be no rationale being used for the choice from among the available methods. Clearly, what must be done for each phylogenetic data set is for the nature of the data to be evaluated, and then for a reasoned choice to be made about which methods might be appropriate for subsequent analysis, along with an assessment of the support for the phylogenetic hypotheses generated. Two recent studies that incorporate a carefully-reasoned evaluation of the data, followed by a series of phylogenetic analyses that are themselves then subjected to re-evaluation, are those of Lento et al. (1995) and Penny et al. (1995), and these could serve as useful models.

Acknowledgements—Thanks to Steve Barker, Peter Baverstock, Dave Booth, Mike Crisp, John Ellis, Dan Faith, Alan Johnson and John Trueman for commenting on an earlier version of the manuscript; and to John Ellis for providing the ssrRNA data file.


Takazaki N. & Nei M. 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *Journal of Molecular Evolution* 38: 210-218.


