# The powers and pitfalls of parsimony

Caro-Beth Stewart

**Parsimony analysis is a powerful tool for the study of biological evolution. It is used to construct phylogenetic trees, to evaluate alternative hypotheses objectively, and to study evolutionary pattern and process. Yet, as comparative data sets expand, the pitfalls of parsimony analysis are catching experts and novices alike.**

THE principle of parsimony states that the simplest explanation consistent with a data set should be chosen over more complex explanations, and is a guiding tenet in scientific study (for critical review, see ref. 1). Parsimonious reasoning is a fundamental way of 'knowing' in comparative evolutionary biology, whether the raw data are molecular, physiological, morphological or behavioural[1-3]. In particular, the principle of parsimony is the foundation of a powerful method, called parsimony analysis, that allows reconstruction of the past to be undertaken with logical and statistical rigour[4-6]. Here I outline the powers and pitfalls of parsimony analysis, with emphasis on its application to molecular sequence data[7-11]. The following represents my opinion as regards a reasonable consensus in the field; it is written with the novice in mind.

Darwin[12] referred to evolution as 'descent with modification', and this simple phrase still embodies our best current understanding of the process[1,4,13]. The goal of building phylogenetic trees is to discover the genealogical relationships between 'taxa' (biological entities such as genes, proteins, individuals, populations, species, or higher taxonomic units). A phylogeny is a hierarchy of nested sets of taxa indicating relative recentness of common ancestry[4-6,12-15].

A widespread misconception, especially concerning molecular sequences[16], is that the building of evolutionary trees simply requires the grouping of taxa according to overall similarity[1,4,13]. Several methods that embody this concept have been invented: they are referred to as distance-matrix methods[9,17-20]. These methods ignore the possibility that apparent overall similarity and true evolutionary relationship are not necessarily the same thing[1,4-6,21]. Two taxa can appear quite similar to each other yet be related relatively distantly. (After all, who among us would assume that an Elvis Presley look-alike is actually his twin brother?) Conversely, two closely related taxa may appear quite different from each other.

This distinction can be illustrated by the genealogical relationship of tetrapods (the four-legged vertebrates) to their fish-like ancestors. In Fig. 1, we consider the possible evolutionary relationships between three vertebrate lineages, one leading to sharks, another to lungfishes, a third to primates. If these species were grouped according to overall morphological similarity, then the tree that unites sharks and lungfishes (Fig. 1, tree 1) would be chosen. Yet, careful morphological comparisons of extant and fossil vertebrates clearly indicates that lungfishes and tetrapods are more closely related to one another than either are to sharks[22] (that is, tree 3 is correct). The apparent similarity between sharks and lungfishes is due to the retention of ancestral characteristics; these two lineages have evolved more slowly than has the lineage leading to primates.

Although molecular sequences generally evolve at a more regular 'clock-like' rate than morphological characters[23], there are many known examples of the same molecule evolving at different rates in different species[24]. For example, baboon $\alpha$-globin differs from rhesus monkey $\alpha$-globin by 9 amino acids and from human $\alpha$-globin by 11 amino acids, whereas the human and rhesus proteins differ by only 5 amino acids[24,25]. Of these three species, the two monkeys are most closely related. Assuming that the three $\alpha$-globins are the products of the 'same' globin gene duplicate in the three primates which diverged when the species diverged (that is, the genes are 'orthologous'[11]), then the two monkey $\alpha$-globins are most closely related; the large difference in their amino-acid sequences is most probably due to very rapid evolution of the baboon protein[24,25]. Yet, if these three primate molecules were grouped by overall similarity, the human and rhesus $\alpha$-globins would cluster together. The above distinction between overall similarity and evolutionary relationship clearly applies with equal force to molecules; therefore, the task of building trees that accurately reflect evolutionary history is often more complicated than simply clustering taxa by overall similarity.

It has long been recognized[14,26] that overall similarity can be broken into three components (Fig. 2): (1) shared-ancestral characters, which are due to retention of traits found in the common ancestor; (2) shared-derived characters, which are due to new traits or modifications that arose along more recent lines of common descent; and (3) homoplasies (convergences, parallelisms and reversals), which are due to the same new trait or modification having been derived independently along different lineages[27]. The concept of overall similarity consciously groups 'true' evolutionary resemblance (shared-ancestral characters and shared-derived characters) together with 'false' resemblance (homoplasy)[4]. Willi Hennig[14] formalized the concept that relative closeness of evolutionary relationship

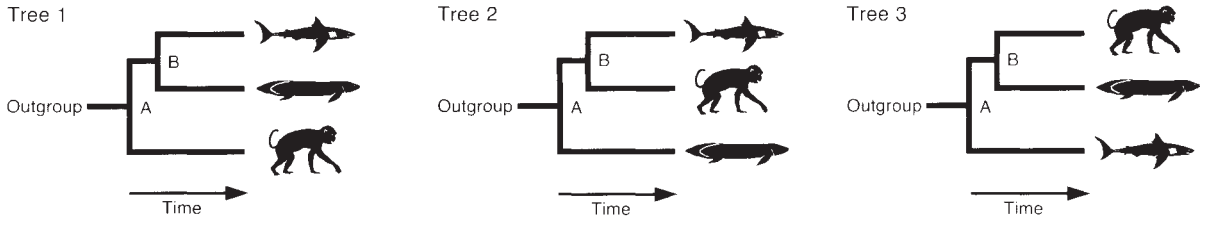



FIG. 1 Three possible trees relating sharks, lungfishes, and tetrapods. Shown here are the three possible bifurcating, rooted (node A) evolutionary trees relating sharks, lungfishes and tetrapods. In tree 1, sharks and lungfishes are depicted as most closely related; that is, they share a more recent common ancestor (node B) than either does with tetrapods (represented here by a monkey). In tree 2, sharks and monkeys are depicted as most closely related. In tree 3, lungfishes and monkeys are depicted as most closely related. Tree 3 is thought to be correct[22].

should be inferred solely on the basis of shared-derived characters. Hennig's ideas spawned the school of thought called cladistics[4-6,15,28] from which modern parsimony analysis developed[7-11,13].

## Powers of parsimony

The fundamental powers of parsimony analysis are that it uses derived characters to infer phylogenetic trees (Box 1), and that it can be used to tweeze apart overall similarity into its various components[14] (Fig. 2). Parsimony analysis may well be the most versatile and powerful tool in evolutionary biology; but, like any power tool, its proper use requires training, practice and attentiveness.

**Building phylogenetic trees.** The essence[18,29] of building parsimony trees from molecular sequences[7-11] is illustrated in Box 1. At its simplest, the method is a purely logical one that partitions similarities on a character-by-character basis. Alternative trees are evaluated, one character at a time, to determine how many evolutionary events they each require. By the criterion of parsimony, the 'best' or most parsimonious tree is the one that requires the fewest total events[30,31]. At the operational level, parsimony analysis does not distinguish between shared-derived similarities and homoplasies[32]. The most parsimonious tree will be the correct phylogeny only if the number of shared-derived characters is high enough and the number of homoplasies low enough. Recent studies using known phylogenies and actual molecular data have shown that parsimony analysis is generally reliable for the inference of the correct tree[33,34]. Parsimony and other methods of phylogenetic tree-building have been reviewed recently[9,17,19].

It is simple to examine all possible trees for four or five taxa manually. Indeed, building trees by hand is the best way to understand the method, and is a highly recommended exercise (Box 1). But, as the number of taxa increases, the number of possible trees increases in a greater than exponential manner[30], and building trees quickly becomes a task best suited for a computer.

A versatile and sophisticated computer package for inferring parsimony trees, *Phylogenetic Analysis Using Parsimony* (PAUP)[32], has been developed by David Swofford. Although other good parsimony programs are available (listed in refs 9 and 20), PAUP is the most generally useful package, and will be discussed here. PAUP can guarantee to find the most parsimonious trees for relatively large data sets, and has many invaluable features and options. Using various user-defined assumptions, PAUP can analyse any type of character data (such as nucleic acid sequences, protein sequences, restriction site polymorphisms, morphological characters, behavioural characters, and so on). The ability to analyse diverse data sets with the same basic method and computer

### BOX 1 Building parsimony trees using nucleotide sequences

The simplest case[18] for undirected parsimony analysis involves a minimum of four taxa. There are three possible unrooted bifurcating trees for four taxa, as illustrated in *a*, using shark (S), lungfish (L), monkey (M) and a non-vertebrate outgroup (O). Parsimony analysis of nucleotide sequences requires homologous (evolutionarily related) and properly aligned sequences; each nucleotide position in the sequence is considered a 'character' which can have five states (G, A, T, C or gap). Ten nucleotides from a hypothetical gene for the four taxa are aligned in *b*. Characters 1, 3 and 7 are invariant (shared-ancestral for all four taxa); invariant and highly conserved characters allow alignments and inference of homology (evolutionary relationship) between sequences. Only 'phylogenetically informative' sites (those that can be used to choose between alternative trees, in this case because two taxa have one nucleotide and two other taxa have another nucleotide) are useful in building parsimony trees[18]. The phylogenetically informative characters are marked with an asterisk (*). The other variable sites (characters 4 and 10) are not phylogenetically informative because all trees require the same number of substitutions.

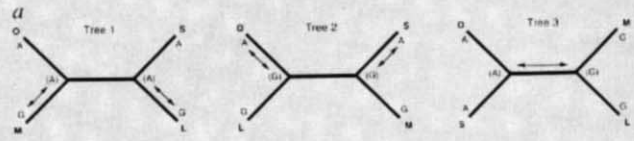
**Choosing between alternative trees:** To find the most parsimonious tree, each character is evaluated on the unrooted trees to determine how few substitutions are needed to explain its observed distribution. For example, character 2 is analysed on the three unrooted trees in *a*. As indicated by the arrows, tree 1 requires a minimum of two nucleotide substitutions (one of two equally parsimonious solutions is shown, with A assumed at the nodes, and changes to G occurring along the lineages leading to M and L), tree 2 requires a minimum of two substitutions (again, only one of the parsimonious solutions is shown), and tree 3 requires only one substitution. If each character in the sequence were analysed in this manner (and the reader should do this), the events per tree would be as shown in *b*.

The optimal or 'best' tree by the parsimony criterion is the one that requires the fewest total evolutionary events[30,31] (nucleotide substitutions, in this case); that tree is said to be the 'most parsimonious' or 'shortest' tree. The shortest tree in this hypothetical example is tree 3, because it requires 10 substitutions, whereas trees 1 and 2 each require 12. The sites that are compatible with each tree are underlined; the number of compatible sites per tree can be used in statistical tests in hopes of ruling out alternatives[29,39].

**Rooting the tree:** The tree produced by parsimony programs such as PAUP[32] is 'unrooted', meaning that the ancestral node has not been identified. Rooting any tree requires knowledge or assumptions about the data set; decisions regarding rooting cannot be abdicated to the computer. Three common ways of rooting molecular trees are as follows. (1) Outgroup. One or more lineages can be included that are known to have diverged before the divergence of the taxa of interest. Ideally, the outgroup taxon should be a member of a closely related sistergroup to the taxa of interest, the ingroup. (2) Gene duplication. When dealing with multigene families, a known earlier gene duplicate can be used as an outgroup to root the tree[10]. If gene duplicates are used, a final step is needed to complete the rooted tree: gene duplications must be assigned at internal nodes in sufficient number to explain the known distribution of the duplicates within extant species. (3) Midpoint. If no information is available regarding outgroups, the root can be placed at the half-way mark, or midpoint, along the longest reconstructed lineage between two taxa[32]. Midpoint rooting rests on the shaky assumptions of relatively equal rates of evolution along different lineages and appropriate sampling of the lineages.

To visualize turning an unrooted tree into a hierarchical phylogeny, think of the unrooted tree as though it were made of string, and mentally tug on the ancestral node until the other lineages move into place (or make a string tree and do this manually). Tugging on the outgroup of each of the unrooted trees shown here will produce the respective rooted trees in Fig. 1.

*a*




*b*

| Taxa | Character | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 5 | | | | | 10 |
| Shark | G | A | T | C | C | T | A | G | G | C |
| Lungfish | G | G | T | C | A | C | A | T | G | T |
| Monkey | G | G | T | C | A | T | A | T | C | T |
| Outgroup | G | A | T | A | C | C | A | G | C | A |
| | | * | | | * | * | | * | * | |
| Events per tree 1 | 0 | 2 | 0 | 1 | 2 | 2 | 0 | 2 | 1 | 2 |
| Events per tree 2 | 0 | 2 | 0 | 1 | 2 | 1 | 0 | 2 | 2 | 2 |
| Events per tree 3 | 0 | 1 | 0 | 1 | 1 | 2 | 0 | 1 | 2 | 2 |

*Total:* 12 (Events per tree 1)
*Total:* 12 (Events per tree 2)
*Total:* 10 (Events per tree 3)

program is helping to unite the subdisciplines of evolutionary biology.

In theory, a tree built from molecular sequences (the 'gene tree') should be an accurate reflection of the evolutionary history of those sequences. An accurate gene tree will mirror the phylogeny of the species from which the molecules were obtained if, and only if, the molecules being compared are orthologous, are not different alleles that have been retained across speciation events, and have not been the victims of gene conversion or other disruptive events after the speciation[18,29,35]. The inference of species phylogenies from gene trees is in widespread practice today, and has yielded many important and unexpected results[36,37]. Gene trees also inform us about the evolution of multigene families[10,18,38].

**Objective evaluation of trees.** Although the statistical properties of phylogenetic trees are not completely understood, several statistical methods have been developed to evaluate tree hypotheses[19,39,40]. One widely used approach is the 'bootstrap' resampling method[19], which is used to evaluate the support for internal branches of a tree; this method is available as an option in PAUP. Another class of methods counts the number of sites compatible with different (usually four-taxa) phylogenies, and applies a statistical test to see if alternative trees can be ruled out[29,39]. The ability to test alternative hypotheses objectively makes phylogenetic reconstruction a 'falsifiable' discipline[40].

**Character and process analysis.** Parsimony analysis attains its greatest power for the study of evolutionary pattern and process when done on rooted trees (Fig. 2), here referred to as cladistic or character analysis. Cladistic analysis can be used to reconstruct character states at the internal and ancestral nodes, and can suggest whether a given character is shared-ancestral, shared-derived, uniquely derived, or a result of homoplasy[27,38,41–46]. This information can be used to infer modes of evolution (what happened), to calculate rates of evolution (how quickly it happened) and to detect adaptation (why it happened)[24,45,46]. Parsimonious reasoning on phylogenies is a rigorous method in comparative biology, as is elegantly explained in two recent books[2,3]. Although this approach is not widely used by molecular biologists, character analysis is a powerful tool for studying the evolution of the molecules themselves[10,11,38,44–46]. For example, the most parsimonious explanation concerning introns in protein-coding genes is that they were inserted during early eukaryotic evolution[47].

Although character analysis can be done manually, *MacClade*[43], a visual and interactive Macintosh computer program, greatly aids in these calculations. The manual to this program provides a helpful review of parsimony and character analysis.

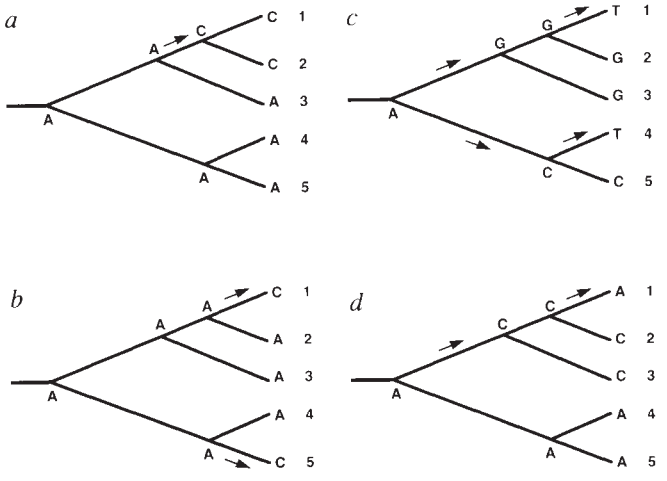



FIG. 2 The different types of similarity. The same phylogenetic tree uniting five taxa (labelled 1–5) is shown in each of the figures. The various reasons why homologous nucleotides from different taxa can be the same[56] are illustrated. These principles also apply to other unpolarized characters (that is, characters such as molecular sequences for which the ancestral state is not known *a priori*). For purposes of illustration, the ancestral states of the characters are assumed to be known. G, guanine; A, adenine; T, thymine; C, cytosine. *a*, Shared-ancestral and shared-derived characters. The Cs observed in taxa 1 and 2 are examples of shared-derived characters, whereas the As in taxa 3, 4 and 5 are shared-ancestral characters. Shared-derived characters indicate relative recentness of common ancestry, whereas shared-ancestral characters do not. Characters that are shared-derived for one level in the hierarchy may be shared-ancestral for more closely related taxa. *b*, Parallel substitutions. The Cs observed in taxa 1 and 5 are the result of nucleotide substitutions that occurred in parallel along those lineages, as indicated by the arrows. If the phylogeny were not known, this site would appear to support a tree that links taxa 1 and 5. By parsimonious reasoning on this known phylogeny, the ancestral states at the internal nodes (that is, A at each node) and the direction of substitution (a change from an ancestral A to a derived C) can be inferred from the observed sequences in the five taxa. *c*, Convergent substitutions. The Ts observed in taxa 1 and 4 are the result of convergent nucleotide substitutions because they arose from different ancestral nucleotides (G and C), and caused the sequences of taxa 1 and 4 to regain sequence similarity over time[26,56]. (In this case, the direction of the substitutions could not be inferred from the observed sequences without further information.) The term convergence is used to mean different things in different contexts[26,27]. For example, unrelated molecules are said to be convergent if they have gained similar functions, regardless of how this was accomplished. In common usage, the term convergence is often used in place of the term homoplasy. *d*, Reversal. The A observed in taxon 1 is the result of a change back to the ancestral state, A, which was retained in taxa 4 and 5. This reversal makes taxon 1 appear more similar to taxa 4 and 5 than it is to its closer relatives, taxa 2 and 3. (If the ancestral states were not known in this case, there would be other equally parsimonious solutions.)

## Pitfalls of parsimony

Parsimony analysis is simple and straightforward: therefore, its problems are relatively well understood. The problems encountered while building parsimony trees generally fall into two broad categories: (1) failure to find the shortest tree, and (2) the shortest tree not being the correct phylogeny. These pitfalls are discussed below, along with possible ways to detect and avoid them.

Many of the pitfalls may be avoided entirely through judicious choices regarding the number and type of taxa and characters collected. Before beginning a study, careful thought should be given to what scientific questions are to be addressed. Regardless of the major scientific questions, certain practical and methodological questions should be answered (see Box 2).

**Failure to find the shortest tree.** The goal of building parsimony trees is to find the shortest tree (or trees) that exists for a given data set under the chosen assumptions[31,32]. This pursuit can fail for two practical reasons: too many taxa or too few phylogenetically informative sites, or a combination of both.

**Too many taxa.** Whenever possible, evolutionary tree programs should be used in a mode that guarantees to find all optimal trees. PAUP[32] has two exact algorithms that guarantee to find all most-parsimonious trees for a limited number of taxa. The task of finding the shortest tree becomes prohibitively time-consuming as the number of taxa increases, even for the most sophisticated and rapid computers and programs. For up to 10 taxa, PAUP can do an exhaustive search of all possible trees, and produce a frequency distribution showing the tree lengths. Using a 'branch and bound' algorithm, PAUP can guarantee to find the shortest tree for a maximum of about 12 to 25 taxa, depending on the data, but will not give the distribution[32]. The practical upper limit for the number of taxa depends on the length and complexity of the character data, the assumptions and options chosen, the speed of the computer, and the patience of the investigator. To analyse a large number of taxa, a computer run may take hours, days or even weeks. Because it often takes longer to analyse data properly than to generate it, phylogenetic

analysis should not be tacked on to the end of a study as an afterthought. A few suggestions concerning the analysis of large data sets follow.

If the data set has too many taxa for exact algorithms, question whether all taxa are necessary. Rarely is the exact placement of every available taxon relevant to the question of interest. Sometimes taxa can be selectively omitted, provided that important conclusions are not altered significantly by the choice of omitted data[48].

Alternatively, a large set of taxa can be broken into smaller groups[49]. For example, one may be interested in the branching order of certain major lineages (such as plants, animals and fungi), for which numerous sequences are available. Rather than omitting taxa, one might find the most parsimonious tree for the species within each major lineage, then define the topologies of these sub-trees to PAUP[32]. PAUP will treat each defined sub-tree as a single taxon while building trees that relate the lineages, thereby greatly reducing the number of alternative trees that must be considered. Moreover, PAUP will reconstruct ancestral sequences for the sub-trees which should better represent the lineages than would any of the individual sequences. Thus, the power of phylogenetic reconstruction[10,11,42,43] can be combined with the hypothesis testing abilities of the four-way test[29,39] to focus on the question of interest.

For some questions, neither of the above taxa reduction approaches may be appropriate. If so, parsimony programs can be run using faster 'heuristic' algorithms that attempt, but do not guarantee, to find the shortest trees. Heuristic parsimony algorithms rearrange starting trees to look for shorter ones, and can get trapped in local optima where no simple rearrangement will yield the globally most parsimonious tree. A way to escape local optima is to use many different starting trees[50]; this can be done automatically in PAUP through its random addition option[32].

Distance trees also can be used as starting trees in heuristic parsimony searches. Currently, most distance tree algorithms are heuristic and will analyse a large number of taxa quickly, yet will produce only one 'optimal' tree per run. With heuristic algorithms, the input order of the data can influence the branching order of the tree that is found. Therefore, the input matrix should be reordered and the program run repeatedly to search for optimal trees[20]. Ideally, optimal distance and parsimony trees should be found and compared; congruence of the topologies is reassuring, but does not guarantee that the correct tree for that data has been inferred. The computer software package, PHYLIP[20], contains programs for most distance-matrix methods, as well as some molecular parsimony programs. Taken together, PAUP and PHYLIP cover most methods for phylogenetic inference.

**Too few phylogenetically informative sites.** If a data set does not contain enough shared-derived, phylogenetically informative characters to resolve the branching order of all of the taxa, many equally parsimonious trees may be found. What constitutes 'enough' such characters depends on the type and quality of data; as a minimum, the data set generally should contain more phylogenetically informative characters than it does taxa. Poorly supported branches in the trees will have low bootstrap scores. Complete lack of resolution of lineages can be detected by branches of zero length. Distance methods sometimes can perform better than parsimony for data sets having few phylogenetically informative sites, because all variable sites are used in the distance calculations[51].

The problem of too few phylogenetically informative characters can arise if too short a segment of DNA is sequenced, or if a region is chosen for study that is not evolving rapidly enough to answer the question at hand. It is a good idea to make analysis an ongoing part of the research strategy, because it can provide valuable information concerning adequacy of the data. The obvious solution to this problem is to collect more appropriate characters per taxa, but this is not always feasible. It may be difficult or impossible to collect enough phylogenetically informative characters to fully resolve a gene phylogeny for very closely related taxa, such as individuals within populations or species.

The data set from the well publicized human mitochondrial DNA study[52] illustrates this point. In this extensive study, 610 nucleotides from the rapidly evolving mitochondrial control region were sequenced from 189 individuals; of the 610 characters, 201 were variable and 119 were phylogenetically informative[52]. This data set contains too many taxa for an exact search, and too few phylogenetically informative sites to allow complete resolution of the phylogeny by parsimony; indeed numerous equally parsimonious trees exist for these data[50,52]. This example should be a cautionary tale for those who wish to study population genetics phylogenetically: the combined problems of too many taxa and too few phylogenetically informative sites are likely to plague all such studies. Indeed, the literature contains numerous examples that have not been so well publicized.

**The shortest tree not being the correct phylogeny.** Even if the data set appears to have plenty of phylogenetically informative sites and an exact search finds one most parsimonious tree, that tree may not accurately reflect the true evolutionary history of the taxa. There are conditions under which parsimony analysis can fail to find the correct phylogenetic tree.

Homoplasy, in its various disguises, is the ultimate trickster of parsimony. Identities or similarities due to convergence, parallel evolution, and reversals can cause historically incorrect trees to be most parsimonious. Convergence and parallel evolution are often considered indications of adaptation or positive selection, but a certain amount of homoplasy happens by chance[27,53].

**Chance homoplasy.** The more distantly related the taxa, the more likely that multiple substitutions have occurred at variable sites in the sequences, especially at synonymous sites in codons. Multiple hits obscure the phylogenetic 'signal' (the informative sites that support the true phylogeny) with 'noise' (homoplasy). A noisy data set cannot produce the correct phylogeny with any certainty[54]; yet low levels of random homoplasy are unlikely to produce an incorrect tree having significant support because the events are usually scattered over the tree and cancel each other[53].

**BOX 2 Practical and methodological questions concerning evolutionary trees**

Below are examples of some basic questions that should be answered regarding phylogenetic analyses; generally this information should be presented with published phylogenies. If space permits, the complete sequence alignment should also be presented, as should any derived distance matrices used in the analyses. The primary data ought to be easily available to reviewers and readers who wish to verify results or try additional analyses; submitting the data on diskettes with the manuscript is recommended.

- What computer program and algorithm was used to build the tree(s)?
- Was the program used in a manner that guarantees to find the 'best' tree?
- If not, what measures were taken to find the best tree? (For example, how many times was the data set reordered and the program rerun?)
- What criterion was used to select the tree or trees presented?
- Were there other trees that tied for best? (If so, how many? What do they look like; that is, what are their evolutionary implications?)
- If many equally parsimonious trees were found, does it make more sense to present a consensus tree?
- How many evolutionary events (that is, nucleotide substitutions, amino-acid replacements and so on) does the best tree require?
- Are there other trees that require only a few more events?
- Can these alternative trees be ruled out statistically?
- How much support (that is, bootstrap value) is there for any given branch?
- Is the gene tree a reasonable species tree? If not, what are the possible explanations?

A quick way to check a data set for phylogenetic signal is to examine the distribution of possible trees for skewness[7,54]. PAUP automatically produces tree distributions during exhaustive searches, and a random sample of possible trees can be produced under other search modes[32,54]. A data set with high phylogenetic signal should have a positive skew, with the shortest tree(s) on a tail of the distribution[7,53].

How might phylogenetic signal be extracted from noisy DNA data? One approach is to weight transversions more heavily than transitions[55], which can be done easily in PAUP[32]. For protein-coding genes, third positions in codons can be omitted, and only the more conservative first and second positions analysed[29]. For distantly related proteins, it may be more productive to analyse the amino-acid sequences[16] with protein parsimony programs such as PROTPARS[20,32].

**Directed homoplasy.** More interesting cases of homoplasy are those due to adaptation rather than chance[27]. In unusual cases, 'convergence' of amino-acid sequences can cause distantly related taxa to be pulled together on amino-acid parsimony trees[29,45,46]. A warning sign for such taxonomically localized homoplasy is the discovery of two or more short trees that dramatically rearrange a lineage, combined with an unusually long reconstructed branch length for the 'misplaced' lineage[45]. Cladistic analysis of the sequences on the 'correct' rooted phylogeny will pinpoint the homoplastic characters[45,46]. Molecular 'convergence' or 'parallelism' generally presents itself at the protein, not DNA, sequence level[24,27,45,46]. This highlights the need for cladistic analysis of protein sequences to study adaptive evolution of protein structure and function[38,45].

Although homoplasy may be a pitfall of parsimony tree building, the detection of convergence and parallel evolution is one of the major powers of parsimony analysis. Character analysis on a rooted phylogeny, regardless of how the phylogeny was constructed, is the only method by which homoplasy can be detected[56].

## Conclusion

Genome evolution is a rich and complex tapestry interwoven with chromosome and gene duplications, gene conversions, mobile genetic elements, allelic diversity, hybridization and, in rare cases, sequence convergence and horizontal gene transfer. Evolutionary trees built from molecular sequences reflect these complex processes. Thus, the failure of a molecular phylogeny to reflect the phylogeny of the species perfectly should not be taken as a failure of parsimony analysis. If a gene tree conflicts with an accepted species tree, one should stop and ponder why. If the same tree is found using other genes from the same species, then the molecules are probably correct about the species phylogeny. If not, the different genes may have different evolutionary histories, which can be reconstructed through careful comparative studies. Comparative analysis of genes and proteins in a phylogenetic framework informs us about molecular evolutionary processes, and sheds light on the evolution of genomes. Until recently, the primary use of parsimony analysis has been largely limited to the study of organismal evolution: its potential to resolve questions about molecular evolution is only now being realized. □

*Caro-Beth Stewart is at the Department of Biological Sciences, State University of New York at Albany, Albany, New York 12222, USA.*


1. Sober, E. *Reconstructing the Past: Parsimony, Evolution, and Inference* (MIT Press, Cambridge, MA, 1988).
2. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, New York, 1991).
3. Brooks, D. R. & McLennan, D. A. *Phylogeny, Ecology, and Behavior* (University of Chicago Press, Chicago, 1991).
4. Eldredge, N. & Cracraft, J. *Phylogenetic Patterns and the Evolutionary Process* (Columbia Univ. Press, New York, 1980).
5. Wiley, E. O. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics* (Wiley, New York, 1981).
6. Hull, D. L. *Science as a Process* (University of Chicago Press, Chicago, 1988).
7. Fitch, W. M. in *Cladistics* (eds Duncan, T. & Stuessy, T. F.) 221–252 (Columbia Univ. Press, New York, 1984).
8. Fitch, W. M. *Am. Nat.* **111,** 223–257 (1977).
9. Swofford, D. L. & Olsen, G. J. in *Molecular Systematics* (eds Hillis, D. M. & Moritz, C.) 411–501 (Sinauer, Sunderland, MA, 1990).
10. Dayhoff, M. O. & Eck, R. V. *Atlas of Protein Sequence and Structure* Vol. 2 (National Biomedical Research Foundation, Silver Spring, MD, 1966).
11. Fitch, W. M. *Syst. Zool.* **19,** 99–113 (1970).
12. Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, London, 1859).
13. Hull, D. L. in *The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis* (eds Fernholm, B., Bremer, K. & Jörnvall, H.) 3–15 (Elsevier, Amsterdam, 1989).
14. Hennig, W. *Phylogenetic Systematics* (translated by Davis, D. D. & Zangerl, R.) (University of Illinois Press, Urbana, 1966).
15. Wiley, E. O., Siegel-Causey, D., Brooks, D. R. & Funk, V. A. *The Compleat Cladist: A Primer of Phylogenetic Procedures* (The University of Kansas Museum of Natural History, Special Publication No. 19, 1991).
16. Doolittle, R. *Of URFs and ORFs: A Primer on how to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA, 1986).
17. Nei, M. in *Phylogenetic Analysis of DNA Sequences* (eds Miyamoto, M. M. & Cracraft, J.) 90–128 (Oxford Univ. Press, New York, 1991).
18. Li, W.-H. & Graur, D. *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA, 1991).
19. Felsenstein, J. *A. Rev. Genet.* **22,** 521–565 (1988).
20. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) Version 3.5* (Computer software package and manual distributed by the author, Dept. Genetics, University of Washington, Seattle, WA, 1993).
21. Gould, S. J. *Nat. Hist.* **92,** 14–21 (1992).
22. Colbert, E. H. & Morales, M. *Evolution of the Vertebrates* 4th edn (Wiley, New York, 1991).
23. Wilson, A. C., Carlson, S. S. & White, T. J. *A. Rev. Biochem.* **46,** 573–639 (1977).
24. Gillespie, J. *The Causes of Molecular Evolution* (Oxford Univ. Press, New York, 1991).
25. Shaw, J.-P., Marks, J., Shen, C. C. & Shen, C.-K. J. *Proc. natn. Acad. Sci. U.S.A.* **86,** 1312–1316 (1989).
26. Haas, O. & Simpson, G. G. *Proc. Am. phil. Soc.* **90,** 319–349 (1946).
27. Patterson, C. *Molec. Biol. Evol.* **5,** 603–625 (1988).
28. Nelson, G. & Platnick, N. *Systematics and Biogeography: Cladistics and Vicariance* (Columbia Univ. Press, New York, 1981).
29. Irwin, D. M. & Wilson, A. C. in *Mammalian Phylogeny* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) 257–276 (Springer, New York, 1992).
30. Cavalli-Sforza, L. L. & Edwards, A. W. F. *Evolution* **32,** 550–570 (1967).
31. Farris, J. S. *Syst. Zool.* **19,** 83–92 (1970).
32. Swofford, D. L. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0 s* (Computer program and manual distributed by the Center for Biodiversity, Illinois Natural History Survey, Champaign, IL 61820, 1992).
33. Atchley, W. R. & Fitch, W. M. *Science* **254,** 554 (1991).
34. Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. & Molineux, I. J. *Science* **255,** 589–592 (1992).
35. Doyle, J. *Syst. Bot.* **17,** 144–163 (1992).
36. Wilson, A. C., Zimmer, E. A., Prager, E. & Kocher, T. D. in *The Hierarchy of Life* (eds Fernholm, B., Bremer, K. & Jörnvall, H.) 407–419 (Elsevier, Amsterdam, 1989).
37. Graur, D., Hide, W. A. & Li, W.-H. *Nature* **351,** 649–652 (1991).
38. Stewart, C.-B. *Meth. Enzym.* (in the press).
39. Li, W.-H. & Gouy, M. *Meth. Enzym.* **183,** 645–659 (1990).
40. Penny, D., Hendy, M. D. & Steel, M. A. in *Phylogenetic Analysis of DNA Sequences* (eds Miyamoto, M. M. & Cracraft, J.) 155–183 (Oxford Univ. Press, New York, 1991).
41. Swofford, D. L. & Maddison, W. P. in *Systematics, Historical Ecology, and North American Freshwater Fishes* (ed. Mayden, R. L.) 186–223 (Stanford Univ. Press, Stanford, 1992).
42. Maddison, W. P. & Maddison, D. R. *Folia Primatol.* **53,** 190–202 (1989).
43. Maddison, W. P. & Maddison, D. R. *MacClade: Analysis of Phylogeny and Character Evolution* (Sinauer, Sunderland, MA, 1992).
44. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. *FEBS Lett.* **262,** 104–106 (1990).
45. Stewart, C.-B., Schilling, J. W. & Wilson, A. C. *Nature* **330,** 401–404 (1987).
46. Swanson, K. W., Irwin, D. M. & Wilson, A. C. *J. molec. Evol.* **33,** 418–425 (1991).
47. Palmer, J. D. & Logsdon, J. M. Jr *Curr. Opin. Genet. Dev.* **1,** 470–477 (1991).
48. Patterson, C. in *The Hierarchy of Life* (eds Fernholm, B., Bremer, K. & Jörnvall, H.) 471–488 (Elsevier, Amsterdam, 1989).
49. Sankoff, D., Cedergren, R. J. & McKay, W. *Nucleic Acids Res.* **10,** 421–431 (1982).
50. Maddison, D. R., Ruvolo, M. & Swofford, D. L. *Syst. Biol.* **41,** 111–124 (1992).
51. Cornish-Bowden, A. *J. theor. Biol.* **101,** 317–319 (1983).
52. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. *Science* **253,** 1503–1507 (1991).
53. Peacock, D. & Boulter, D. *J. molec. Biol.* **95,** 513–527 (1975).
54. Hillis, D. M. in *Phylogenetic Analysis of DNA Sequences* (eds Miyamoto, M. M. & Cracraft, J.) 278–294 (Oxford Univ. Press, New York, 1991).
55. Fitch, W. M. & Ye, J. in *Phylogenetic Analysis of DNA Sequences* (eds Miyamoto, M. M. & Cracraft, J.) 147–154 (Oxford Univ. Press, New York, 1991).
56. Sneath, P. H. A. & Sokal, R. R. *Numerical Taxomomy: the Principles and Practice of Numerical Classification* (Freeman, San Francisco, 1973).


ACKNOWLEDGEMENTS. I thank R. Collura, K. Helm-Bychowski, D. Irwin, W. Messier, L. Taylor and D. Swofford for commenting on various versions of the manuscript; R. Collura for figures; D. Hillis for preprints and discussions; D. Swofford for a test version of PAUP; D. Maddison and W. Maddison for a test version of *MacClade*; and J. Felsenstein for PHYLIP. This paper is dedicated to the memory of Allan C. Wilson.