Phylogeny

We have now seen several examples of evolution, in proteins and in genomes. These represent the extension to the molecular level of concerns that have occupied biologists since Darwin and even before. The basic principle is that *the origin of similarity is common ancestry.* Although there are many exceptions, arising from convergent evolution, the importance of this principle both for rationalizing contemporary observations and giving a window into the history of life cannot be overestimated.

The field of phylogeny has the goals of working out the relationships among species, populations, individuals, or genes.[3] Relationship is taken in the literal sense of kinship or genealogy, that is, assignment of a scheme of descendants of a common ancestor (see Box). The results are usually presented in the form of an evolutionary tree. The taxonomy of the ratites - large flightless birds - is a typical example (Fig. 4.7a). The ancestor of the ratites is believed to be a bird that could fly, probably related to the extant tinamous.

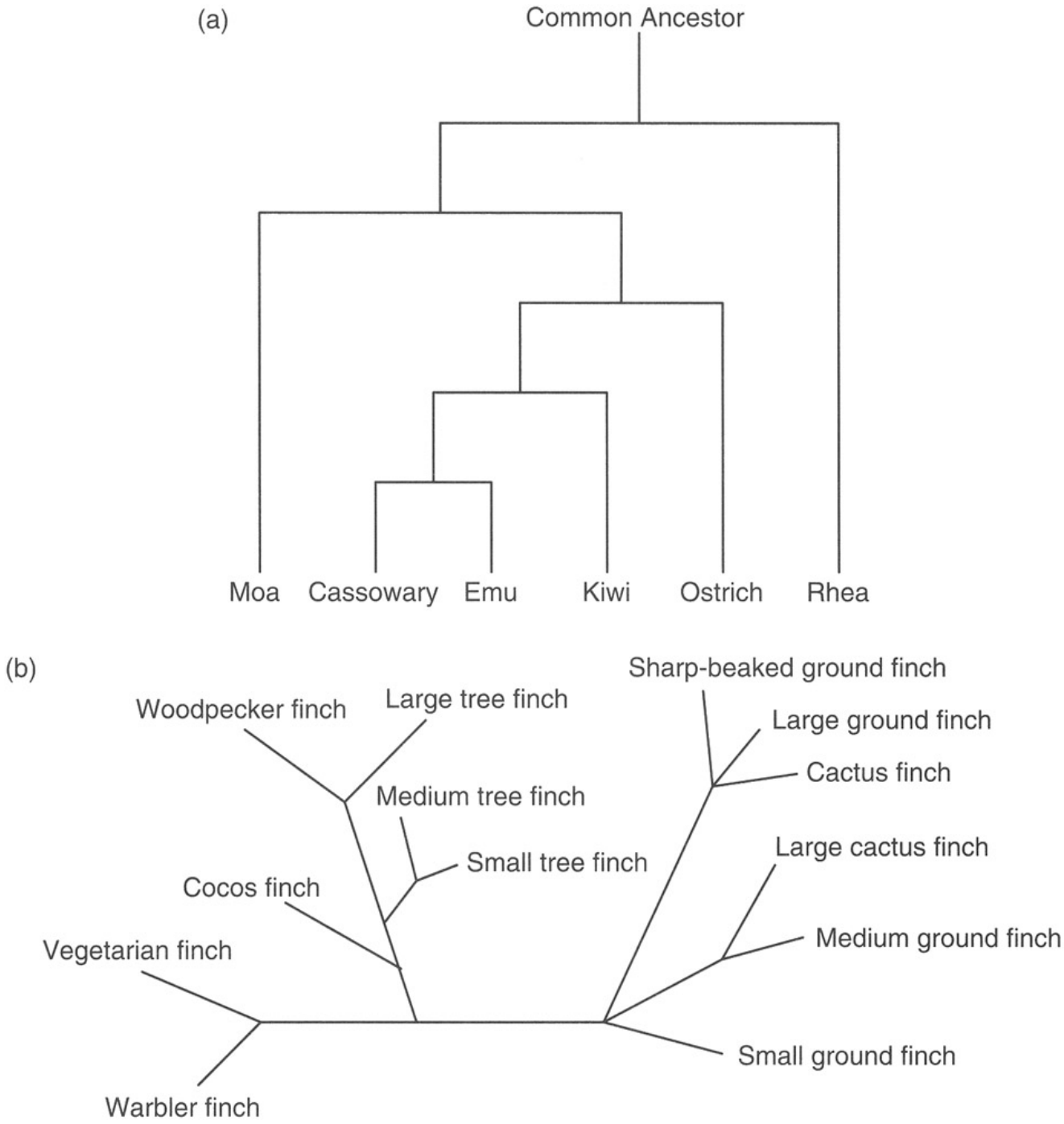


Figure 4.7: (a) Phylogenetic tree of ratites (large flightless birds) based on mitochondrial DNA sequences. The common ancestor is at the *root* of this tree. A surprising implication of these DNA sequences is that the moa and kiwi are not closest relatives, and therefore that New Zealand must have been colonized twice by ratites or their ancestors. (b) *Unrooted* tree of relationships among finches from the Galapagos and Cocos Islands. Darwin studied the Galapagos finches in 1835, noting the differences in the shapes of their beaks and the correlation of beak shape with diet. Finches that eat fruits have beaks like those of parrots, and finches that eat insects have narrow, prying beaks. These observations were seminal to the development of Darwin's ideas. As early as 1839 he wrote, in *The Voyage of the Beagle*, 'Seeing this gradation and diversity of structure in one small, intimately related group of birds, one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends.'

Concepts related to biological classification and phylogeny

Homology means, specifically, descent from a common ancestor.
Similarity is the measurement of resemblance or difference, Independent of the source of the resemblance. Similarity is observable in data collectable *now,* and involves no historical hypotheses. In contrast, assertions of homology require inferences about historical events which are almost always unobservable.

Clustering is bringing together similar items, distinguishing classes of objects that are more similar to one another than they are to other objects outside the classes. Most people would agree about degrees of similarity, but clustering is more subjective. When classifying objects, some people prefer larger classes, tolerating wider variation; others prefer smaller, tighter, classes. They are called *groupers* or *splitters.*
Hierarchical clustering is the formation of clusters of clusters of...
Phylogeny is the description of biological relationships, usually expressed as a tree. A statement of phylogeny among objects *assumes* homology and *depends* on classification. Phylogeny states a topology of the relationships based on classification according to similarity of one or more sets of characters, or on a model of evolutionary processes. In many cases, phylogenetic relationships based on different characters are consistent, and support one another. If different characters induce inconsistent phylogenetic relationships, they are all dubious. Conversely, note that the same similarity data may be consistent with different possible topologies or trees.

Such a tree, showing all descendants of a single original ancestral species, is said to be *rooted.* (The root of the tree typically appears at the top or the side; botanists will have to get used to this.) Alternatively, we may be able to specify relationships but not order them according to a history. The relationships among the finches of the Galapagos Islands, studied by Darwin, plus a related species from the nearby Cocos Island, are shown in an *unrooted tree* (Fig. 4.7b). Addition of data from a species on the South American mainland ancestral to the island finches would allow us to *root the tree.*

Statement of a tree of relationships may reveal only the connectivity or topology of the tree, in which case the lengths of the branches contain no information. A more ambitious goal is to show the distances between taxa quantitatively; for instance, to label the branches with the time since divergence from a common ancestor. Given a set of data that characterize different groups of organisms - for example, DNA or protein sequences, or protein structures, or shapes of teeth from different species of animals - how can we derive information about the relationships among the organisms in which they were observed? It is rare for species relationships and ancestry to be directly observable. Evolutionary trees determined from genetic data are often based on inferences from the patterns of similarity, which are all that is observable among species living now. We generally assume that the more similar the characters the more closely related the species, although this is a dangerous assumption. Nevertheless, from the relationships among the characters we wish to infer patterns of ancestry: the *topology* of the phylogenetic relationships (informally, the 'family tree.')
To what extent do the topologies of the relationships depend on the choice of character? In particular, are there *systematic* discrepancies between the implications of molecular and palaeontological analysis? Molecular approaches to phylogeny developed against a background of traditional taxonomy, based on a variety of morphological characters, embryology, and, for fossils, information about the geological context (stratigraphy). The classical methods have some advantages. Traditional taxonomists have much less restricted access to extinct organisms, via the fossil record. They can *date* appearances and extinctions of species by geological methods. Molecular biologists, in contrast, have very limited access to extinct species. Some sub-fossil remains of species which became extinct as recently as the last century or two have legible DNA, including specimens of the quagga (a relative of the zebra) and the thylacine (Tasmanian 'wolf', a marsupial), and some New Zealand birds (including moas). We have already seen an example of a sequence from the mammoth. Some DNA sequences from Neanderthal man have been recovered from an individual who died approximately 30 000 years ago. But *Jurassic Park* remains fiction!

A crucial event in the acceptance of molecular methods occurred in 1967 when V.M. Sarich and A.C. Wilson dated the time of divergence of humans from chimpanzees at 5 million years ago, based on immunological data. At that time palaeontologists dated this split at 15 million years ago, and were reluctant to accept the molecular approach. Reinterpretation of the fossil record led to acceptance of a more recent split, and broke the barrier to general acceptance of molecular methods.

Indeed, many molecular properties have been used for phylogenetic studies, some surprisingly long ago. Serological cross-reactivity was used from the beginning of the last century until superseded by direct use of sequences. In one of the most premature scientific studies I know of, E.T. Reichert and A.P. Brown published, almost a century ago (in 1909), a phylogenetic analysis of fishes based on haemoglobin crystals. Their work was based on Stenö's law (1669), that although different crystals of the same substance have different dimensions - some are big, some small - they have the same interfacial angles, reflecting the similarity in microscopic arrangement and packing of the atomic or molecular units within the crystals. Reichert and Brown showed that the interfacial angles of crystals of haemoglobins isolated from different species showed patterns of similarity and divergence parallel to the species' taxonomic relationships.

Reichert and Brown's results are replete with significant implications. They show that proteins have definite, fixed shapes, an idea by no means recognized at the time. They imply that as species progressively diverge, the structures of their haemoglobins progressively diverge also. In 1909, no one had a clue about nucleic acid

or protein sequences. In principle, therefore, the recognition of evolution of protein structures preceded, by several decades, the idea of evolution of sequences.

Today, DNA sequences provide the best measures of similarities among species for phylogenetic analysis. The data are digital. It is even possible to distinguish selective from non-selective genetic change, using the third position in codons, or untranslated regions such as pseudogenes, or the ratio of synonymous to non-synonymous codon substitutions. Many genes are available for comparison. This is fortunate, because, given a set of species to be studied, it is necessary to find genes that vary at an appropriate rate. Genes that remain almost constant among the species of interest provide no discrimination of degrees of similarity. Genes that vary too much cannot be aligned. There is an analogous situation in radioactive dating requiring choice of an isotope with a half-life of the same general magnitude as the time interval to be determined.

Fortunately, genes vary widely in their rates of change. The mammalian mitochondrial genome, a circular double-stranded DNA molecule approximately 16 000 by long, provides a useful fast-changing set of sequences for study of evolution among closely-related species. In contrast, ribosomal RNA sequences were used by C. Woese to identify the three major divisions: Archaea, Bacteria and Eukarya.

Conversely, different rates of change of sequences of different genes can lead to different and even contradictory results in phylogenetic studies. This is especially true if what we want is not just the topology of the relationships but the branch lengths. In addition, horizontal gene transfer, and convergent evolution, are competing phenomena - that is, competing with descent - that interfere with the deduction of phylogenetic relationships.

[3]The general term is 'taxa.' The *observable* taxa - for instance the extant species for which we wish to work out the pattern of ancestry, are called the 'operational taxonomic units', abbreviated to OTUs.

Phylogenetic trees

We describe phylogenetic relationships as trees. In computer science, a tree is a particular kind of graph. A graph is a structure containing nodes (abstract points) connected by edges (represented as lines between the points) (see Box). A *path* from one node to another is a consecutive set of edges beginning at one point and ending at the other, like our trip from Malmö to Tromsø. A *connected graph* is a graph containing at least one path between any two nodes. From these we can define a *tree:* a connected graph in which there is *exactly* one path between every two points. A particular node may be selected as a *root*; but this is not necessary - abstract trees may be rooted or unrooted (see Fig. 4.7). Unrooted trees show the topology of relationship but not the pattern of descent. A rooted tree in which every node has two descendants is called a *binary tree* (see PERL program, page 202).
Another special kind of graph is a *directed graph* in which each edge is a one-way street. Examples include the Hidden Markov Model diagram shown in Fig. 4.6, and the neural networks illustrated in Chapter 5. Rooted phylogenetic trees are, implicitly, directed graphs, the ancestor-descendent relationship implying the direction of each edge.

Glossary of terms related to graphs

Graph an abstract structure containing *nodes* (points) and *edges* (lines connecting points).
Path a consecutive set of edges.
Connected graph a graph in which there is at least one path between every two nodes.
Tree a connected graph with exactly one path between every two points.
Edge length a number assigned to each edge signifying in some sense the distance between the nodes connected by the edge.
Path length the sum of the lengths of the edges that comprise the path.

It may be possible to assign numbers to the edges of a graph to signify, in some sense, a 'distance' between the nodes connected by the edges. The graph may then be drawn to scale, with the sizes of the edges proportional to the assigned lengths. The length of a path through the graph is the sum of the edge lengths.
In phylogenetic trees, edge lengths signify either some measure of the dissimilarity between two species, or the length of time since their separation. The assumption that differences between properties of living species reflects their divergence times will be true only if the rates of divergence are the same in all branches of the tree. Many exceptions are known; for instance, among mammals rodents show relatively fast evolutionary rates for many proteins (see Weblem 4.8).
Broadly, there are two approaches to deriving phylogenetic trees. One approach makes no reference to any historical model of the relationships. Proceed by measuring a set of distances between species, and generate

the tree by a hierarchical clustering procedure. This is called the *phenetic* approach. The alternative, the *cladistic* approach, is to consider possible pathways of evolution, infer the features of the ancestor at each node, and choose an optimal tree according to some model of evolutionary change. Phenetics is based on similarity; cladistics is based on genealogy.

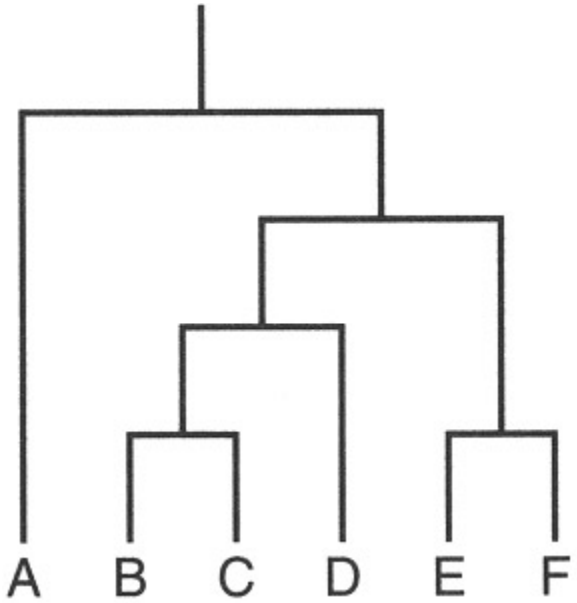
Clustering methods

Phenetic or clustering approaches to determination of phylogenetic relationships are explicitly non-historical. Indeed, hierarchical clustering is perfectly capable of producing a tree even in the absence of evolutionary relationships. A departmental store has goods clustered into sections according to the type of product - for instance, clothing or furniture - and subclustered into more closely related sub-departments, such as men's and women's shoes. Men's and women's shoes have a common ancestor, but there is no implication that shoes and furniture do.

```
#!/usr/bin/perl

#drawtree.prl --  draws binary trees (root at top)

#usage:  echo '(A((BC)D)(EF))'  | drawtree.prl > output.ps



print <<EOF;

%!PS-Adobe-\n%%BoundingBox: atend

/n /newpath load def /m /moveto load def /l /lineto load def

/rm /rmoveto load def /rl /rlineto load def /s /stroke load def

1.0 setlinewidth 50 100 translate 2 2 scale

/Helvetica findfont 10 scalefont setfont

EOF



$tree = <>; chop($tree); $_ = reverse($tree); s/[()]//g;



$x = 0; $y = 0;

while ($nd = chop()) {

    print "$x $y m ($nd) stringwidth pop -0.5 mul 0 rm ($nd) show\n";

    $xx{$nd} = $x; $x+=20; $yy{$nd) = 10;

}



while ($tree =~ s/\(?([A-Z])([A-Z])\/)?\/1/) {

    print "n $xx{$l) $yy{$1) m\n";

    ($yy{$1} > $yy{$2}) || {$yy{$1} = $yy{$2}}; $yy{$l} += 20;

    print "$xx{$1} $yy{$1} 1 $xx{$2} $yy{$1) 1 $xx{$2} $yy{$2} l s\n";
```

```
    $xx{$1} = 0.5*($xx{$l) + $xx{$2});
}
print "n $xx{$tree} $yy{$tree} m 0 20 rl s showpage\n";

$rx = 2*$x + 30; $yt = 2*$yy{$tree} + 146;
print "%%BoundingBox: 40 95 $rx $yt\n";
```

A PERL program to draw binary trees. The input: (A((BC)D)(EF)) produces the following output, as a PostScript file, which can be printed on most printers and displayed on most terminals.



A simple clustering procedure works as follows: given a set of species, determine for all pairs a measure of the similarity or difference between them. This could depend on a physical body trait such as the difference between the average adult height of members of two species. Or one could use the number of different bases in alignments of mitochondrial DNA. To create a tree from the set of dissimilarities, first choose the two most closely related species and insert a node to represent their common ancestor. Then replace the two selected species by a set containing both, and replace the distances from the pair to the others by the average of the distances of the two selected species to the others. Now we have a set of pairwise dissimilarities, not between individual species, but between sets of species. (Regard each remaining individual species as a set containing only one element.) Then repeat the process, as in the following example.

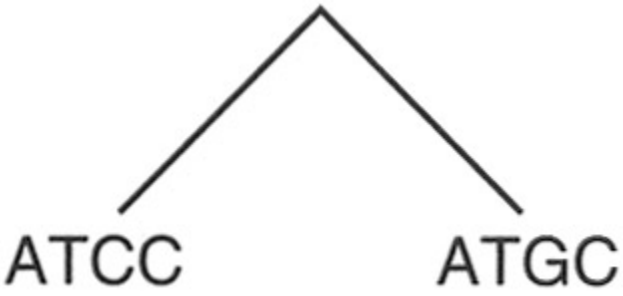
Example 4.7

Consider four species characterized by homologous sequences ATCC, ATGC, TTCG, and TCGG. Taking the number of differences as the measure of dissimilarity between each pair of species, use a simple clustering procedure to derive a phylogenetic tree.

The distance matrix is:

	ATCC	**ATGC**	**TTCG**	**TCGG**
ATCC	0	**1**	2	4
ATGC		0	3	3
TTCG			0	2
TCGG				0

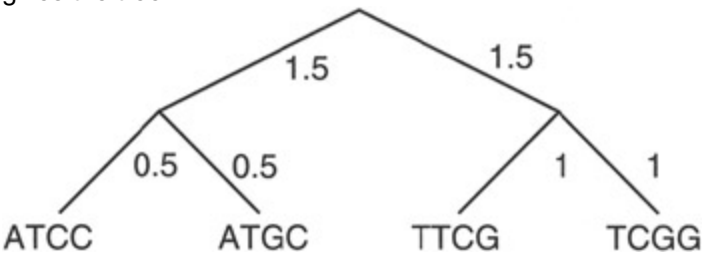
Because the matrix is symmetric, we need fill in only the upper half. The smallest distance is 1 (in boldface), between ATCC and ATCG. Therefore, our first cluster is {ATCC, ATGC}. The tree will contain the fragment:



The reduced distance matrix is:

	{ATCC, ATGC}	**TTCG**	**TCGG**
{ATCC, ATGC}	0	1/2 (2 + 3) = 2.5	1/2z (4 + 3) = 3.5
TTCG		0	**2**
TCGG			0

The next cluster is {TTCG, TCGG}, distance **2**. Finally, linking the clusters {ATCC, ATGC} and {TTCG, TCGG} gives the tree:



Branch lengths have been assigned according to the rule: branch length of edge between nodes *X* and *Y* = ½ distance between *X* and *Y*

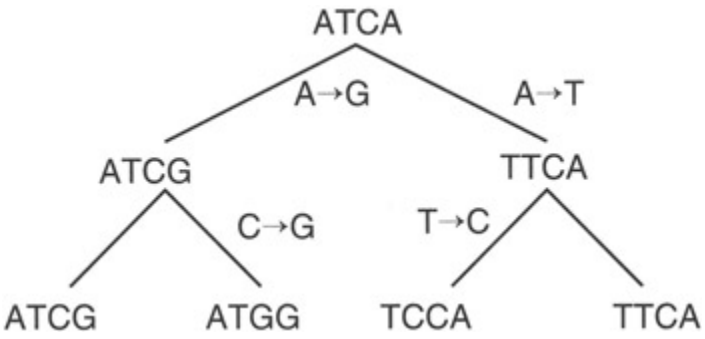
Whether the branch lengths are truly proportional to the divergence times of the taxa represented by the nodes must be determined from external evidence.

This process of tree building is called the UPGMA method (Unweighted Pair Group Method with Arithmetic mean). A modification of the UPGMA method by N. Saitou and M. Nei, called Neighbour Joining, is designed to correct for unequal rates of evolution in different branches of the tree.

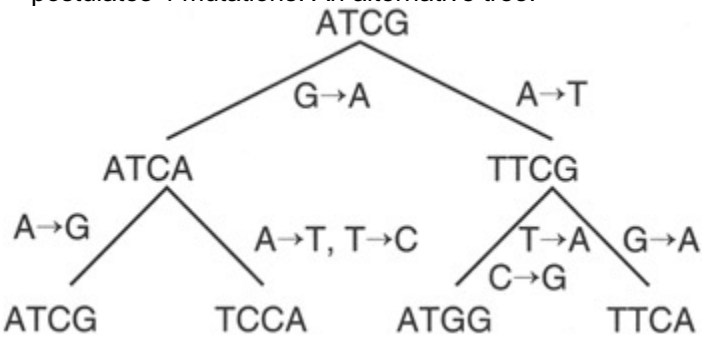
Cladistic methods

Cladistic methods deal explicitly with the patterns of ancestry implied by the possible trees relating a set of taxa. Their aim is to select the correct tree by utilizing an explicit model of the evolutionary process. The most popular cladistic methods in molecular phylogeny are the *maximum parsimony* and *maximum likelihood* approaches. They are specialized to sequence data, starting from a multiple sequence alignment. Neither maximum parsimony nor maximum likelihood could be applied to anatomic characters such as average adult height.

The *maximum parsimony* method of W. Fitch defines an optimal tree as the one that postulates the fewest mutations. For instance, given species characterized by homologous sequences ATCG, ATGG, TCCA, and TTCA, the tree:



postulates 4 mutations. An alternative tree:



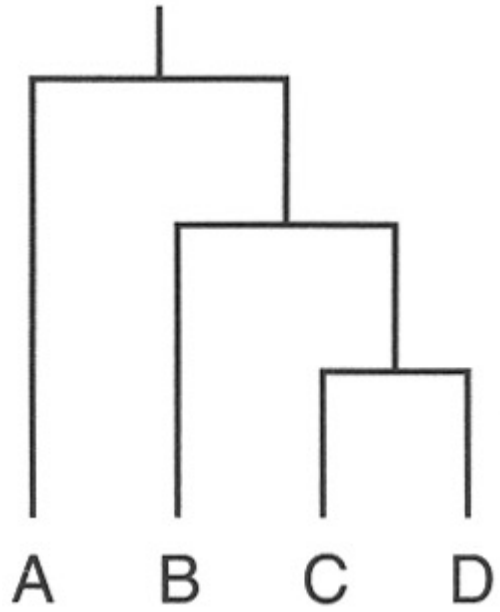
postulates 7 mutations. Note that the second tree implies that the G→A mutation in the fourth position occurred twice independently. The former tree is optimal according to the maximum parsimony method, because no other tree involves fewer mutations. In many cases, several trees may postulate the same number of mutations, fewer than any other tree. For such cases the maximum parsimony approach does not give a unique answer.

The *maximum likelihood* method assigns quantitative probabilities to mutational events, rather than merely counting them. Like maximum parsimony, maximum likelihood reconstructs ancestors at all nodes of each tree considered; but it also assigns branch lengths based on the probabilities of the mutational events postulated. For each possible tree topology, the assumed substitution rates are varied to find the parameters that give the highest likelihood of producing the observed sequences. The optimal tree is the one with the highest likelihood of generating the observed data.

Both maximum parsimony and maximum likelihood methods are superior to clustering techniques. This has been demonstrated with cases where independent evidence - for instance, from classical palaeontology - provides a correct answer, and also with simulated data - computed generation of evolving sequences.

The problem of varying rates of evolution

Suppose that the four species A, B, C, and D have the phylogenetic tree:



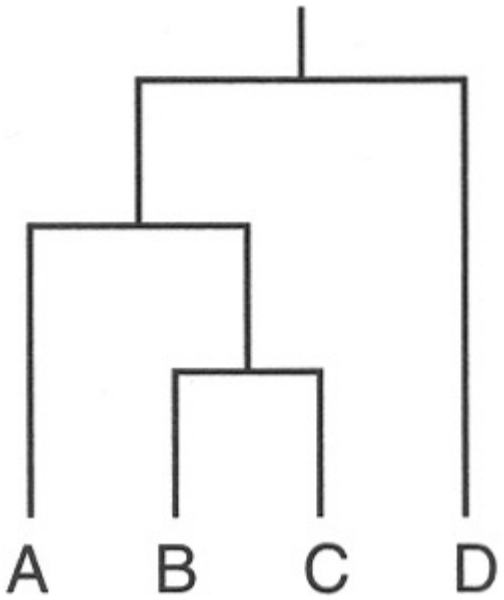
This tree is consistent with the dissimilarity matrix:

	A	B	C	D
A	0	3	3	3
B		0	2	2
C			0	1
D				0

Suppose however that species D is changing very fast, although the phylogeny is unaltered. The dissimilarity matrix might then be observed to be:

	A	B	C	D
A	0	3	3	20
B		0	2	20
C			0	20
D				0

from which we would derive the incorrect phylogenetic tree:



All the methods discussed here are subject to errors of this kind if the rates of evolutionary change vary along different branches of the tree. To test for varying rates, compare the species under consideration with an *outgroup - a* species more distantly related to all the species in question than any pair of them is to each other. For instance, if we are studying species of primates, a non-primate mammal such as the cow would be a suitable outgroup. If the rates of evolution among the primate species were constant, we should expect to observe approximately equal dissimilarity measures between all primate species and the cow. If this is not observed, the suggestion is that evolutionary rates have varied among the primates, and the character being used may well not provide the correct phylogenetic tree.

Computational considerations

Cladistic methods - maximum parsimony and maximum likelihood - are more accurate than simpler clustering methods such as UPGMA, but require large amounts of computer time if the number of species is appreciable. The total number of possible trees, which cladistic methods are committed to considering if they could, increases very rapidly with the number of species. As a result, in many cases of interest these methods can give only approximate answers, even with respect to their intrinsic assumptions.

Because calculated phylogenies are often approximations, it is important to try to test them. Methods include:

1. Comparison of phylogenies obtained from different characters describing the same set of taxa - are they consistent? If trees produced from different characters share a subtree, perhaps that portion of the phylogeny has been determined reliably and other portions have not.
2. Analysis of subsets of taxa should give the same answer - with respect to the subset - as appears within the full tree.
3. Formal statistical tests, involving re-running the calculation on subsets of the original data, are known as *jackknifing* and *bootstrapping:*
   - *Jackknifing* is calculation with data sets sampled randomly from the original data. For phylogeny calculations from multiple sequence alignments, select different subsets of the positions in the alignment, and rerun the calculation. Finding that each subset gives the same phylogenetic tree lends it credibility. If each subset gives a different tree, none of them is trustworthy.
   - *Bootstrapping* is similar to jackknifing except that the positions chosen at random may include multiple copies of the same position, to form data sets of the same size as the original, to preserve statistical properties of the sampling.
4. If there are very long edges, consider seriously the possibility of unequal variation in evolutionary rate that may have perturbed the calculation. Introduce outgroup taxa to check.

Web Resource: Phylogenetic Trees

The taxonomic community has expended great effort to produce mature software. The PHYLIP package (PHYLogeny Inference Package) of J. Felsenstein is an integrated collection of many different techniques. The programs work on many different types of computers, and are freely distributed and easily obtained.

Summaries of tools for phylogenetics; includes useful list of web sites, and general listing of phylogeny software: http://evolution.genetics.washington.edu/phylip/software.html and Whelan, S., Liò, P. and Goldman, N. (2001) Molecular phylogenetics: State-of-the-art methods for looking into the past, *Trends in Genetics* 17, 262–272.

Some multiple sequence alignment packages, such as CLUSTAL-W, provide facilities to launch a phylogenetic tree calculation from the alignments they produce.

Recommended reading

Altschul, S.F. and Koonin, E.V. (1998) 'Iterated profile searches with PSI-BLAST — a tool for discovery in protein databases', *Trends in Biochemical Sciences* 23, 444–7. [Description of one of the most important tools for database searching for sequence similarity.]

Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. (1994) 'Issues in searching molecular sequence databases', *Nature Genetics* 6, 119–29. [General background to challenges in designing information retrieval methods and interpreting the results.]

Eddy, S. (1996) 'Hidden Markov models', *Current Opinion in Structural Biology* 6, 361–5. [Readable introduction to an important mathematical technique providing powerful tools for detection of distantly-related sequences, and protein fold recognition.]

Efron, B. and Gong, G. (1983) 'A leisurely look at the boostrap, the jackknife, and cross-validation', *The American Statistician* 37, 36–48. [Classic paper on statistical methods for calibrating pattern recognition procedures.]

Li, W-H. (1997) *Molecular Evolution* (Sunderland, MA, USA: Sinauer.) [A detailed discussion of evolution and phylogenetic analysis.]

Penny, D., Hendy, M.D., Zimmer, E.A., and Hamby, R.K. (1990) 'Trees from sequences: Panacea or Pandora's box?', *Australian Systematic Botany* 3, 21–38. [Cautionary notes about determination of phylogenetic trees.]