

# Molecular Phylogeny Reconstruction

Sudhir Kumar, *Arizona State University, Tempe, Arizona, USA*

Alan J Filipski, *Arizona State University, Tempe, Arizona, USA*

Molecular phylogenetics deals with the inference of evolutionary relationships among individuals, populations, species and higher taxonomic entities using molecular data.

## Introduction

In the second half of the twentieth century many laboratory techniques became available for examining diversity within and among species by analysis of biologically important molecules. These include methods based on cross-reactivity of antibodies, protein electrophoresis, DNA–DNA (deoxyribonucleic acid) hybridization, restriction fragment length polymorphism, and direct sequencing of DNA and proteins (polypeptides). Of these, comparisons of DNA sequences are the most informative and powerful. Within the last decade, complete DNA sequences of many genomes have been obtained and the public sequence repositories are bulging with sequence information for thousands of genes from diverse species.

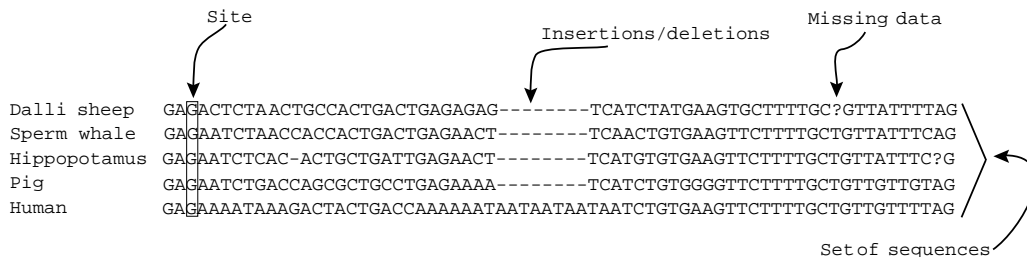
In comparison to the classical approach to phylogeny reconstruction, each site (numbered position) in a set of homologous sequences can be considered as corresponding to a character, and the identity of the nucleotide base at that site corresponds to the state of that character for the given sequence (Figure 1). Use of molecular sequence data has several advantages compared to the morphological characters used traditionally. For instance, no subjective appraisal is involved in the determination of character state; laboratory techniques tell us exactly which nucleotide base is present at a site. Another advantage is that the same set of states (four bases) applies to all organisms, and thus we can directly compare the most diverse life forms. In addition, the amount of available data is enormous, and it is relatively simple to obtain pertinent data for a given set of

species in the laboratory. Concurrent availability of low-cost, powerful computers and new software algorithms has led to the routine use of molecular sequences in reconstructing evolutionary histories of organisms at various taxonomic levels. In the following, we discuss the methods of molecular phylogenetic reconstruction for DNA because DNA sequences are used most widely. These discussions also hold true for protein sequence data. Detailed account of methods for these and other types of data can be found elsewhere (Nei, 1987; Hillis *et al.*, 1996; Nei and Kumar, 2000).

## Methods

### Assembling a DNA sequence dataset

To infer the evolutionary relationship of a set of organisms using molecular sequence data, we must first ensure that the sequences being compared are homologous. We begin by selecting a gene with a homologue in each organism under study. In fact, the chosen sequences should not only be homologous, but should satisfy the stronger condition of being orthologous, i.e. having diverged by speciation events rather than by gene duplications. Researchers determine sequence orthology by using criteria such as the overall sequence identity and functional similarity, and through the analysis of multigene families to which the sequences belong. One must then decide whether to use the nucleotide sequence of the gene or the amino acid sequence



**Figure 1** An alignment of a portion of the  $\gamma$ -fibrinogen gene sequence from five mammals. Insertion–deletion mutations predicted by sequence alignment are shown with hyphens (-) and the missing data is shown with question marks (?).

Secondary article

Article Contents

- Introduction
- Methods
- Impact on Phylogenetics
- Variable Rates

of its protein product, if any. There are many considerations involved. For distantly related organisms, amino acid sequences are often used, because nucleotide sequences evolve much faster than amino acid sequences owing to the redundancy of the genetic code. On the other hand, nucleotide sequences can be more informative, for example, by allowing a distinction to be made between nucleotide substitutions that do not alter the amino acid encoded (silent substitutions) and those that do (replacement substitutions). For intraspecific population genetic studies and for closely related interspecies studies, mitochondrial DNA is often used because parts of it evolve more rapidly than nuclear genes and thus provide more variation for reconstructing evolutionary history.

## Sequence alignment

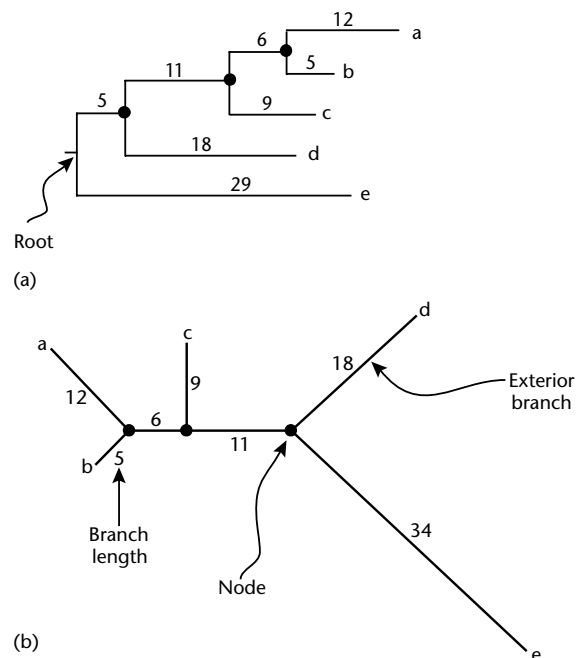
The next step is to align the corresponding positions in different sequences. This is not trivial because sequences of a given gene often differ in length in different species as a result of insertion and deletion mutations. Many computational algorithms and tools are available for this purpose (e.g. Higgins *et al.*, 1996). They work by inserting place holder symbols, usually hyphens, in the sequences to maximize the similarity at each site, while minimizing the cost associated with the number of place holders inserted (Figure 1). At this point, the sequences are of the same length and can be organized into columns, each representing a homologous site. Alignment of DNA sequences from distantly related species or fast evolving genes is generally more difficult. For this reason, DNA sequences that code for protein products are aligned by first constructing an alignment of corresponding amino acid sequences. The protein sequence alignment is then used as a guide to obtain the alignment of the underlying DNA sequences.

## Inferring the phylogenetic tree

At this point, we are ready to infer the phylogenetic tree. The essential structure of the tree is given by its topology, i.e. which nodes are connected to which others (Figure 2). Almost all methods for reconstructing phylogenetic trees produce unrooted trees (Figure 2b). In this case, one may 'root' the tree using a known outgroup sequence. In addition to the branching pattern, we are usually also interested in the length of the branches. Elucidation of the branching pattern is more difficult than the estimation of the branch lengths. In fact, once the topology has been established, one can use statistical methods based on least squares or maximum likelihood approaches for determining the branch lengths (reviewed in Nei and Kumar, 2000).

Several different methods are available for reconstructing phylogenetic trees. Most of them use some criterion (optimality principle) for evaluating the fit of a given dataset to the topology and then search for the tree that

gives the best score in terms of that criterion (see below). If the criterion used is realistic and the data are sufficient, the tree should represent the true phylogenetic relationship of the sequences (and thus the associated organisms). In practical situations, however, this is complicated by the fact that the number of different tree topologies that can be made from a set of sequences increases very rapidly with the number of sequences (Table 1), and we must use heuristics to constrain the search to find a potentially optimal tree quickly. Fortunately, the quality of phylogenetic trees produced by quick heuristics is similar to that obtained with extensive (or exhaustive) searches (Nei and Kumar, 2000).



**Figure 2** Rooted (a) and unrooted (b) tree of five sequences. Branch lengths are drawn proportional to evolutionary distance, which can be expressed in the units of time or the number of substitutions.

**Table 1** Number of possible unrooted trees for different numbers of sequences

Sequences	Trees
3	1
4	3
5	15
6	105
7	945
8	10 395
9	135 135
10	2 027 025
11	654 729 075

At present, three commonly used tree-building criteria in molecular phylogenetics are maximum parsimony (MP), maximum likelihood (ML) and minimum evolution (ME). In the MP criterion, the topology requiring the smallest number of nucleotide changes to fit the observed sequence data is chosen to represent the true tree (Fitch, 1971). In ML methods, the topology with the greatest ML under a given probabilistic model of nucleotide substitutions is chosen (Felsenstein, 1981). In the ME methods, the sequence data are first transformed into a matrix of distances for each sequence pair. Then, the sum of branch lengths needed to fit this matrix to each possible topology is computed, and the topology requiring the smallest sum of branch lengths is chosen. The evolutionary distance between a pair of sequences can be estimated in a number of ways. The simplest distance measure between two sequences is the  $p$ -distance, which is the fraction of sites at which the two sequences differ.  $p$ -distance is known to underestimate the evolutionary distance because of multiple substitutions at the same site. This problem can be remedied by using an appropriate model of nucleotide substitution. A detailed explanation for estimating distances under different models of substitutions and guidelines on choosing appropriate distance measures can be found in Nei and Kumar (2000).

The neighbour-joining method (Saitou and Nei, 1987), because of its computational efficiency, has become frequently used in molecular phylogenetics, especially for large-scale data analyses. It is based on the ME criterion. The neighbour-joining method works in a stepwise fashion by minimizing the sum of branch lengths at each step of sequence clustering. Unweighted pair-group method using arithmetic averages (UPGMA) is another distance-matrix based method, in which pairs of sequences showing the smallest evolutionary distance are clustered first. This method assumes that the evolutionary rate has remained constant throughout the evolutionary history of the given set of organisms. Since this assumption is rarely met in reality, UPGMA should be used cautiously for inferring phylogenetic histories. Many academic software packages are available for computing distances and inferring phylogenetic trees, e.g. Kumar *et al.* (2001).

The choice of which tree-building method to use is somewhat arbitrary and often depends on time requirements, or the philosophical predisposition of the researcher. This is because: (1) no method is uniformly better in reconstructing the true tree when the sequence length is small; and (2) all methods tend to perform well given enough data (Nei and Kumar, 2000).

## Assessing reliability

The next step in constructing a sequence phylogeny is to assess the reliability of the inferred branching pattern. This is often accomplished by a bootstrap analysis (Felsenstein,

1985). Bootstrap procedures involve construction of new sequence sets by resampling with replacement sites (columns) of the original set, building a tree for each new set, and calculating the percentage of times a cluster reappears in the bootstrap replications. This percentage is called the bootstrap value; clusters with a bootstrap value  $\geq 95\%$  are widely considered to reflect correct relationships, although some authors have suggested that 70% may be a more realistic cutoff point. For a detailed explanation of the bootstrap test and information on other types of tests of phylogenetic trees, see Nei and Kumar (2000).

## Impact on Phylogenetics

In general, molecular phylogenetics studies have supported traditional phylogenies constructed on the basis of nonmolecular characters and provided resolution to debates. However, there have been some considerable disagreements between molecular and classical phylogenies. One such example involves the identification of the closest living relatives of the hippopotamus (family Hippopotamidae). Before molecular phylogenetic analyses, Hippopotamidae was thought to be most closely related to Suina (pigs and peccaries), within the mammalian order Artiodactyla. Molecular studies using mitochondrial and nuclear DNA sequences have now clearly established that Hippopotamidae is a sister group to Cetacea (containing whales and dolphins), and that this group is more closely related to ruminants (cows, sheep and deer) than to pigs and peccaries (e.g. Gatesy, 1997). Molecular phylogenetics has also resolved the human–chimpanzee–gorilla trichotomy, identified Chimpanzee *Simian immunodeficiency virus* as the closest relative of the *Human immunodeficiency virus type 1*, and provided insights into sister group relationships of animals and fungi.

## Variable Rates

Molecular phylogenetics would be simpler if all sites in a gene evolved at the same rate (uniform substitution rate among sites) and if all species evolved at the same rate in a given gene (uniform evolutionary rate among lineages). Within a gene, however, certain sites are under stronger natural selection than others because of their functional importance. This variability in evolutionary rates among sites is often accounted for in phylogenetic inference by using a gamma model of nucleotide substitution (Yang, 1996).

## Variable evolutionary rates among lineages

The observed heterogeneity of evolutionary rates among lineages in a gene is caused partly by the nondeterministic nature of the evolutionary processes and partly by differences in intensity and type of natural selection. Tree-building methods mentioned above, with the exception of UPGMA, do not assume constancy of evolutionary rate among lineages (molecular clock) and can thus be used directly. However, some methods are known to produce consistently incorrect results when the evolutionary rates vary significantly among lineages. For instance, Felsenstein (1988) showed that if a four-sequence tree contains two long and two short branches, then the long branches tend to cluster together in the MP trees even if they are distantly related (long-branch attraction problem). One way to avoid this problem is by using a larger number of sequences such that the long branches are broken. Another way to avoid long-branch attraction is to use ML or ME methods.

In summary, molecular phylogenetics has become an integral part of research endeavours in diverse areas of molecular biology, population genetics, developmental biology and evolutionary biology.

## References

- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**: 368–376.
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.
- Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Reviews in Genetics* **22**: 521–565.
- Fitch W (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**: 406–416.
- Gatesy JC (1997) More DNA support for the Cetacea/Hippopotamidae clade: the blood-clotting protein gene gamma-fibrinogen. *Molecular Biology and Evolution* **14**: 537–543.
- Higgins DG, Thompson JD and Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* **266**: 383–402.
- Hillis DM, Moritz C and Mable BK (1996) *Molecular Systematics*. Sunderland, MA: Sinauer.
- Kumar S, Tamura K, Jakobsen I and Nei M (2001) *MEGA: Molecular Evolutionary Genetics Analysis*. [http://www.megasoftware.net]
- Nei M (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press.
- Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.
- Saitou N and Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **6**: 514–525.
- Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* **11**: 367–371.

## Further Reading

- Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis*. Cambridge: Cambridge University Press.
- Graur D and Li W-H (1999) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer.
- Li W-H (1997) *Molecular Evolution*. Sunderland, MA: Sinauer.
- Page RDM and Holmes EC (1998) *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science.
- Patthy L (1999) *Protein Evolution*. Oxford: Blackwell Science.