

Bioinformatics explained: phylogenetics

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history i.e. their *phylogeny*. Phylogenetics is therefore an integral part of the science of *systematics* that aims to establish the phylogeny of organisms based on their characteristics. Furthermore, phylogenetics is central to evolutionary biology as a whole as it is the condensation of the overall paradigm of how life arose and developed on earth.

The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 1 shows a proposed phylogeny for the great apes, *Hominidae*, taken in part from Purvis [Purvis, 1995]. The tree consists of a number of nodes (also termed vertices) and branches (also termed edges). These nodes can represent either an individual, a species, or a higher grouping and are thus broadly termed taxonomical units. In this case, the terminal nodes (also called leaves or tips of the tree) represent extant species of *Hominidae* and are the *operational taxonomical units* (OTUs). The internal nodes, which here represent extinct common ancestors of the great apes, are termed *hypothetical taxonomical units* since they are not directly observable.

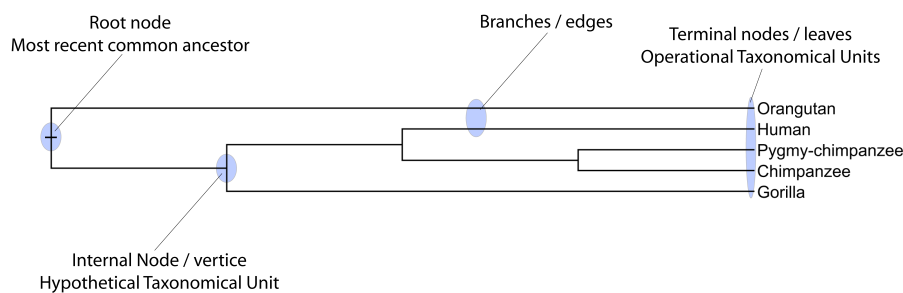


Figure 1: A proposed phylogeny of the great apes (*Hominidae*). Different components of the tree are marked, see text for description.

The ordering of the nodes determine the tree *topology* and describes how lineages have diverged over the course of evolution. The branches of the tree represent the amount of evolutionary divergence between two nodes in the tree and can be based on different measurements. A tree is completely specified by its topology and the set of all edge lengths.

The phylogenetic tree in figure 1 is rooted at the most recent common ancestor of all *Hominidae* species, and therefore represents a hypothesis of the direction of evolution e.g. that the common ancestor of gorilla, chimpanzee and man existed before the common ancestor of chimpanzee and man. If this information is absent trees can be drawn as unrooted.

Modern usage of phylogenies

Besides evolutionary biology and systematics the inference of phylogenies is central to other areas of research.

As more and more genetic diversity is being revealed through the completion of multiple genomes, an active area of research within bioinformatics is the development of comparative machine learning algorithms that can simultaneously process data from multiple species [Siepel

and Haussler, 2004]. Through the comparative approach, valuable evolutionary information can be obtained about which amino acid substitutions are functionally tolerant to the organism and which are not. This information can be used to identify substitutions that affect protein function and stability, and is of major importance to the study of proteins [Knudsen and Miyamoto, 2001]. Knowledge of the underlying phylogeny is, however, paramount to comparative methods of inference as the phylogeny describes the underlying correlation from shared history that exists between data from different species.

In molecular epidemiology of infectious diseases, phylogenetic inference is also an important tool. The very fast substitution rate of microorganisms, especially the RNA viruses, means that these show substantial genetic divergence over the time-scale of months and years. Therefore, the phylogenetic relationship between the pathogens from individuals in an epidemic can be resolved and contribute valuable epidemiological information about transmission chains and epidemiologically significant events [Leitner and Albert, 1999], [Forsberg et al., 2001].

Reconstructing phylogenies from molecular data

Traditionally, phylogenies have been constructed from morphological data, but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as *molecular phylogeny*. The data is most commonly represented in the form of DNA or protein sequences, but can also be in the form of e.g. restriction fragment length polymorphism (RFLP).

Methods for constructing molecular phylogenies can be distance based or character based.

Distance based methods

Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences. To correct for multiple substitutions it is common to use distances corrected by a model of molecular evolution such as the Jukes-Cantor model [Jukes and Cantor, 1969].

UPGMA. A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA) ([Michener and Sokal, 1957], [Sneath and Sokal, 1973]). This method works by initially having all sequences in separate clusters and continuously joining these. The tree is constructed by considering all initial clusters as leaf nodes in the tree, and each time two clusters are joined, a node is added to the tree as the parent of the two chosen nodes. The clusters to be joined are chosen as those with minimal pairwise distance. The branch lengths are set corresponding to the distance between clusters, which is calculated as the average distance between pairs of sequences in each cluster.

The algorithm assumes that the distance data has the so-called *molecular clock* property i.e. the divergence of sequences occur at the same constant rate at all parts of the tree. This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure.

Neighbor Joining. The neighbor joining algorithm, [Saitou and Nei, 1987], on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages, i.e., the tree does not have a particular root. Some programs always draw trees with roots for practical reasons, but for neighbor joining trees, no particular biological hypothesis is postulated by the placement of the root. The method works very much like UPGMA. The main difference is that instead of

[Yang and Rannala, 1997].

Interpreting phylogenies

Bootstrap values

A popular way of evaluating the reliability of an inferred phylogenetic tree is bootstrap analysis. The first step in a bootstrap analysis is to re-sample the alignment columns with replacement. I.e., in the re-sampled alignment, a given column in the original alignment may occur two or more times, while some columns may not be represented in the new alignment at all. The re-sampled alignment represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree.

If a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, this increases the confidence in the original tree. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable. By re-sampling a number of times it is possible to put reliability weights on each internal branch of the inferred tree. If the data was bootstrapped a 100 times, a bootstrap score of 100 means that the corresponding branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability.

Other useful resources

The Tree of Life web-project

<http://tolweb.org>

Joseph Felsensteins list of phylogeny software

<http://evolution.genetics.washington.edu/phylip/software.html>

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

References

- [Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376.
- [Forsberg et al., 2001] Forsberg, R., Oleksiewicz, M. B., Petersen, A. M., Hein, J., Bøtner, A., and Storgaard, T. (2001). A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology*, 289(2):174–179.
- [Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). *Mammalian Protein Metabolism* (ed. HN Munro), chapter Evolution of protein molecules, pages 21–32. New York: Academic Press.
- [Knudsen and Miyamoto, 2001] Knudsen, B. and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A*, 98(25):14512–14517.
- [Larget and Simon, 1999] Larget, B. and Simon, D. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol*, 16:750–759.
- [Leitner and Albert, 1999] Leitner, T. and Albert, J. (1999). The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A*, 96(19):10752–10757.
- [Michener and Sokal, 1957] Michener, C. and Sokal, R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11:130–162.
- [Purvis, 1995] Purvis, A. (1995). A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci*, 348(1326):405–421.
- [Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.
- [Siepel and Haussler, 2004] Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol*, 11(2-3):413–428.
- [Sneath and Sokal, 1973] Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.
- [Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, 14(7):717–724.