Survey of molecular phylogenetics

M. Talianová

Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

ABSTRACT

Rapidly increasing amount of biological data necessarily requires techniques that would enable to extract the information hidden in the data. Methods of molecular phylogenetics are commonly used tools as well as objects of continuous research within many fields, such as evolutionary biology, systematics, epidemiology, genomics, etc. The evolutionary process not only determines relationships among species, but also allows prediction of structural, physiological and biochemical properties of biomolecules. The article provides the reader with a brief overview of common methods that are currently employed in the field of molecular phylogenetics.

Keywords: evolutionary model; distance-based methods; maximum parsimony; maximum likelihood; Bayesian inference; accuracy of phylogeny

Biological sequences (DNA, RNA and amino acids) are complex sources of genetic variation due to various mechanisms such as local changes in DNA sequences, rearrangements of DNA segments or DNA acquisition by horizontal gene transfer (reviewed in Arber 2000). Thus, the comparative analyses of genes and whole genomes enable an exciting view into evolutionary processes and relationships between genetic materials of different living organisms. The evolutionary process not only determines relationships among species, but also allows prediction of structural, physiological, and biochemical properties (Chambers et al. 2000).

Phylogenetic construction is a hierarchical process

Molecular phylogenetics is a continuously evolving area, using and developing methods that enable to extract necessary information. Most of the techniques used in phylogenetic analyses produce phylogenetic trees (phylogenies), which represent evolutionary histories of compared species. Reconstruction of molecular phylogenetic relationships using DNA, RNA or amino acid sequences is a hierarchical process consisting of four steps: (1) alignment of homological sequences, (2) selection of an appropriate mathematical model describing sequence evolution, (3) application of a suitable tree-building method with regard to the analysed data, and (4) assessment of the quality of the resulting phylogeny and interpretation of obtained results (Steel 2005).

Data and models of sequence evolution

Molecular phylogenetics can utilize various characters, such as genome-level characters (Boore 2006) (e.g. position of mobile genetic elements, genome re-arrangements, gene order position, etc.), but it mostly analyses data in the form of biomolecular sequences (nucleic acids or amino acids). Sequences for phylogenetic study are either generated in laboratory or retrieved from sequence databases and aligned. Correct alignment of sequences is a fundamental prerequisite for phylogenetic relationship reconstruction (Harrison and Langdale 2006). Each of the sequences is a subject of random (stochastic) influence of very complex evolutionary processes. Although often very sim-

Supported by the Czech Science Foundation, Projects Nos. 204/05/P505 and 521/05/2076, and by the Institutional Research Plan, Project No. AV0Z50040507.

This contribution was presented at the 4th Conference on Methods in Experimental Biology held in 2006 in the Horizont Hotel in Šumava Mts., Czech Republic.

plified, evolutionary processes can be described using mathematical models of evolution. Some models have very simple assumptions, while others are very complex with numerous parameters representing various biologically relevant facts of sequence evolution. Examples of such parameters are branch lengths of trees (interspeciation times and rates of mutation along the branches), parameters associated with the substitution matrix (e.g. transition/transversion bias), or parameters that describe how mutation rates vary across sites in the sequence. The knowledge of the nature of data used in analysis is an important assumption when choosing a model of evolution. The most of the tree-building methods require mathematical model of sequence evolution, to either compute "distances" between sequences (number of differences corrected for backward, parallel or multiple substitutions) or to explicitly evaluate the probabilities of changes between characters (nucleotides or amino acids) in all positions in the sequence. The simplest is the Jukes-Cantor model (Jukes and Cantor 1969) assuming equal frequency of nucleotides and equal substitution rates. More realistic models are HKY model (Hasegawa et al. 1985), General reversible model (GTR) (Rodríguez et al. 1990), Gamma-distributed-rates models (Wakeley 1994, Yang 1994) and Covarion models (Tuffley and Steel 1998). Considering evolution on the protein level, commonly used models are Dayhoff model of protein evolution (Dayhoff et al. 1978), JTT models (Jones et al. 1992), Codon mutation model (Goldman and Yang 1994), VT model (Muller and Vingron 2000), WAG model (Whelan and Goldman 2001) and many others.

The selection and assessment of the most suitable model is a crucial issue in the phylogenetic reconstruction. To statistically test the accuracy of mathematical models various methods have been developed. It is possible to perform a comparison of two models using likelihood ratio tests (LRTs) (LRT can be performed only for testing nested models, where one model is a special case of the other) (e.g. Huelsenbeck and Crandall 1997), Akaike information criterion (AIC) (Akaike 1974) or Bayesian information criterium (BIC) (Schwarz 1974); (e.g. Huelsenbeck and Crandall 1997), Akaike information criterion (AIC) (Akaike 1974) or Bayesian information criterium (BIC) (Schwarz 1974); otherwise, it is possible to test the overall adequacy of a particular model using parametric bootstrapping (e.g. Whelan et al. 2001) or Bayesian posterior prediction (Huelsenbeck et al. 2001).

Tree-building methods can be classified according to several criteria (Hershkovitz and Leipe 1998). The first way is to define them as either algorithm-based or criterion-based. Algorithm-based methods produce a tree by following a series of steps (e.g. clustering algorithms), while criterion-based methods use an optimality criterion (e.g. the least number of changes in the tree or the topology with a greatest probability of giving rise of analysed data) for comparing alternative phylogenies to one another and deciding, which one fits better. The second group of method-classification is represented by distance-based methods versus character-based methods. Distance-based methods compute pairwise distances according to some measure. Then, the actual data are omitted and the fixed distances are used in the construction of trees. Trees derived using character-based methods are optimised according to the distribution of actual data patterns in relation to a specified character.

Distance-based methods require evolutionary distance (i.e. the number of changes that have occurred along the branches between two sequences) between all pairs of taxa. To obtain relatively unbiased estimate of the evolutionary distance, it is useful to apply a specific evolutionary model that makes assumption about the nature of the evolutionary changes. The examples of distance-based methods used in molecular phylogenetics are the **Least-square method** (Cavalli-Sforza and Edwards 1967, Fitch and Margoliash 1967) or the **Unweighted pair-group method using arithmetic averages – UPGMA** (Sokal and Michener 1958). However, the most popular distance-based technique is the **Neighbor-joining method** (Saitou and Nei 1987) based on agglomerative clustering. Its major strength is the substantial computational speed that makes this method suitable for large datasets; the weakness of this method is the loss of sequence information when converting the data to pairwise distances. It also produces only one tree and thus it is not possible to examine competing hypotheses about the relationship between sequences.

Character-based (discrete) methods operate directly on the aligned sequences rather than on pairwise distances. **Maximum parsimony** (Edwards and Cavalli-Sforza 1963, Fitch 1977) does not require an explicit model of sequence evolution (in contrast to neighbor joining or maximum likelihood method); it identifies the tree (or trees) that involves the smallest number of mutational changes (i.e. the shortest tree length or fewest evolutionary steps) necessary to explain the differences among the data under investigation. In many cases, MP methods are superior to other techniques because they are relatively free of assumptions considering nucleotide and amino acid substitution. MP works well when compared sequences are not too divergent, when the rate of nucleotide substitution is relatively constant and the number of nucleotides examined is large. Furthermore, the parsimony analysis is very useful for some types of molecular data (e.g. insertion sequences, insertions/deletions, gene order or short interspersed nuclear elements – SINEs). The typical problem of MP trees is so called "long-branch attraction" (Hendy and Penny 1989) (or similarly "short-branch attraction"). This phenomenon occurs, when rapidly (slowly) evolving sequences are artefactually inferred to be closely related.

Maximum likelihood method (Cavalli-Sforza and Edwards 1967, Felsenstein 1981) requires a stochastic model of sequence evolution over time. The principle of the likelihood is that the explanation, which makes the observed outcome the most likely (i.e. the most probable) to occur, is the one to be preferred. In maximum likelihood, the topology that gives the highest maximum likelihood value is chosen as the final tree. One of the strengths of the maximum likelihood method is the ease with which hypotheses about evolutionary relationships can be formulated. It enables incorporation of complex models to consider biologically important facts of sequence evolution. On the other side, this method is computationally very intensive, and thus its use can be limited for very large datasets.

Recently, likelihood-based **Bayesian inference** using Markov chain Monte Carlo technique (Rannala and Yang 1996) has become a popular and very useful method; it has been applied to numerous problems in evolutionary or systematic biology.

Phylogenetic networks (e.g. Maddison 1997, Huson and Bryan 2006, Jin et al. 2006) enable to model evolutionary processes of organisms where non-tree events (reticulations) took part. The reticulations arise due to horizontal gene transfer, hybrid speciation or recombination events, and thus create specific links among organisms.

Accuracy of phylogenetic tree

With the increasing emphasis in biology on reconstruction of phylogenetic trees, questions have arisen as to how confident one should be in a given phylogenetic tree and how the support for phylogenetic trees should be measured. The most commonly used methods are non-parametric bootstrap test (Felsenstein 1985) and jack-knife test (Efron 1982), based on random resampling of the original dataset (Efron 1982). These techniques provide a measure of "confidence" for each clade of an observed tree, based on the proportion of bootstrap trees showing the same branching pattern. Another way of testing the reliability of phylogeny is parametric Bayesian inference (reviewed in Huelsenbeck et al. 2001) where the parameters such as the tree topology, branch lengths, or substitution parameters, are assessed by posterior probabilities.

However, when assessing accuracy of resulting phylogeny, one should be cautious when interpreting the results. Besides relying on test values, various biologically relevant facts causing artefactual relationships in the phylogeny (e.g. bad experiment design, characteristics of the data, sources of homoplasy – parallelism, convergence, horizontal gene transfer) should be considered.

Implementation of phylogenetic methods

On the website http://evolution.genetics.washington.edu/phylip/software.html#methods is a comprehensive overview of various phylogenetic packages and programs. These are arranged according to different criteria, some of them are free, some commercial.

REFERENCES

Akaike H. (1974): A new look at the statistical model identification. IEEE T. Automat. Contr., *19*: 716–723.

Arber W. (2000): Genetic variation: molecular mechanisms and impact on microbial evolution. FEMS Microbiol. Rev., *24*: 1–7.

Boore J.L. (2006): The use of genome-level characters for phylogenetic reconstruction. Trends Ecol. Evol., *21*: 439–446.

Cavalli-Sforza L.L., Edwards A.W.F. (1967): Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet., *19*: 233–257.

Chambers J.K., Macdonald L.E., Sarau H.M., Ames R.S., Freeman K., Foley J.J., Zhu Y., McLaughlin M.M., Murdock P., McMillan L., Trill J., Swift A., Aiyar N., Taylor P., Vawter L., Naheed S., Szekeres P., Hervieu G., Scott C., Watson J.M., Murphy A., Duzic E., Klein C., Bergsma D.J., Wilson S., Livi P. (2000): A G protein-coupled receptor for UDP-glucose. J. Biol. Chem., *15*: 10767–10771.

Dayhoff M.O., Schwartz R.M., Orcutt B.C. (1978): A model of evolutionary change in proteins. In: Dayhoff M.O. (eds.): Atlas of Protein Sequences and Structure. National Biomedical Research Foundation, Silver Spring, MD.

Edwards A.W.F., Cavalli-Sforza L.L. (1963): The reconstruction of evolution. Heredity, *18*: 553.

Efron B. (1982): The Jackknife, the Bootstrap and other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia.

Felsenstein J. (1981): Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol., *17*: 368–376.

Felsenstein J. (1985): Confidence limits on phylogenies: An approach using the bootstrap. Evolution, *39*: 783–791.

Fitch W.M. (1977): On the problem of discovering the most parsimonious tree. Am. Nat., *111*: 223–257.

Fitch W.M., Margoliash E. (1967): Construction of phylogenetic trees. Science, *155*: 279–284.

Goldman N., Yang Z. (1994): A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol., *11*: 725–736.

Harisson C.J., Langdale J.A. (2006): A step by step guide to phylogeny reconstruction. Plant J., *45*: 561–572.

Hasegawa M., Kishino H., Yano T. (1985): Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol., *22*: 160–174.

Hendy M.D., Penny D. (1989): A framework for the quantitative study of evolutionary trees. Syst. Zool., *38*: 297–309.

Hershkovitz M.A., Leipe D.D. (1998): Phylogenetic analysis. In: Baxevanis A.D., Ouellette B.F.F. (eds.): Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley Interscience, NY.

Huelsenbeck J.P., Crandall K.A. (1997): Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst., *28*: 437–466.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. (2001): Bayesian inference of phylogeny and its impact on evolutionary biology. Science, *294*: 2310–2314.

Huson D.H., Bryant D. (2006): Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol., *23*: 254–267.

Jin G., Nakhleh L., Snir S., Tuller T. (2006): Maximum likelihood of phylogenetic networks. Bioinformatics, *22*: 2604–2611.

Jones D.T., Taylor W.R., Thornton J.M. (1992): The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci., *8*: 275–282.

Jukes T.H., Cantor C.R. (1969): Evolution of protein molecules. In: Munro H.N. (eds.): Mammalian Protein Metabolism. Academic, NY.

Maddison W.P. (1997): Gene trees in species trees. Syst. Biol., *46*: 523–536.

Muller T., Vingron M. (2000): Modeling amino acid replacement. J. Comput. Biol., *7*: 761–776.

Rannala B., Yang Z. (1996): Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol., *43*: 304–311.

Rodríguez F., Oliver J.L., Marin A., Medina R. (1990): The general stochastic model of nucleotide substitution. J. Theor. Biol., *142*: 485–501.

Saitou N., Nei M. (1987): The Neighbor-Joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol., *4*: 406–425.

Schwarz G. (1974): Estimating the dimension of a model. Ann. Stat., *6*: 461–464.

Sokal R.R., Michener C.D. (1958): A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull., *28*: 1409–1438.

Steel M. (2005): Should phylogenetic models be trying to "fit an elephant"? Trends Genet., *21*: 307–309.

Tuffley C., Steel M.A. (1998): Modelling the covarion hypothesis of nucleotide substitution. Math. Biosci., *147*: 63–91.

Wakeley J. (1994): Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol., *11*: 436–442.

Whelan S., Goldman N. (2001): A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol., *18*: 691–699.

Whelan S., Lio P., Goldman N. (2001): Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet., *17*: 262–272.

Yang Z. (1994): Estimating the pattern of nucleotide substitution. J. Mol. Evol., *39*: 105–111.

Received on February 27, 2007

Corresponding author:

Mgr. Martina Talianová, Akademie věd České republiky, Biofyzikální ústav, v. v. i., Královopolská 135, 612 65 Brno, Česká republika
phone: + 420 541 517 247, fax: + 420 541 240 500, e-mail: talianka18@ibp.cz