A short guide to phylogeny reconstruction

E. Michu

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

ABSTRACT

This review is a short introduction to phylogenetic analysis. Phylogenetic analysis allows comprehensive understanding of the origin and evolution of species. Generally, it is possible to construct the phylogenetic trees according to different features and characters (e.g. morphological and anatomical characters, RAPD patterns, FISH patterns, sequences of DNA/RNA, and amino acid sequences). The DNA sequences are preferable for phylogenetic analyses of closely related species. On the other hand, the amino acid sequences are used for phylogenetic analyses of more distant relationships. The sequences can be analysed using many computer programs. The methods most often used for phylogenetic analyses are neighbor-joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference.

Keywords: sequence alignment; phylogenetic analysis; neighbor-joining; maximum parsimony; maximum likelihood; Bayesian inference

Phylogenetic analysis allows comprehensive understanding of the origin and evolution of species. The main aim of this article is to provide basic information and discuss difficulties in the phylogeny reconstruction. The result of phylogeny reconstruction is a phylogenetic tree, which can be either rooted or unrooted. In the unrooted tree, groupings are inferred, but no direction to evolutionary change is implied. It only displays the relationships between the taxons (Figure 1). Unlike the unrooted tree, the rooted tree implies directionality in time and shows the relationships with regard to an outgroup (Figure 2) (reviewed in Doyle and Gaut 2000). To root the tree, it is necessary to add an outgroup, which is a (unrelated) group of species or single species that is not included in the group of species under the study (reviewed in Harrison and Langdale 2006). The outgroup can be selected on the basis of prior knowledge of the group of interest, or may become apparent during the sequence alignment. Generally, the most informative outgroup is the actual sister group.

The first step in the phylogeny reconstruction is to choose species used in the phylogenetic analyses. The selection of species is very important because "wrong" (incongruent) selection (e.g. only few taxa or the "wrong" taxa) can negatively influence the phylogeny reconstruction; an example of incongruence coming from the "wrong" taxa selection is discussed by Soltis et al. (2004).

Generally, it is possible to construct the phylogenetic trees according to different features and characters (e.g. morphological and anatomical characters, RAPD patterns, FISH patterns, sequences of DNA/RNA, and amino acid sequences).

In the case of molecular phylogeny based on sequencing data, another important consideration in building molecular trees from protein-coding genes is, whether to analyse the sequences at the DNA or the protein level. In DNA sequences, there are only four possible nucleotides and provided DNA substitution rates are high, the probability that two lineages will independently evolve the same nucleotide at the same site increases (reviewed in Bergsten 2005). However, the increased number of characters in nucleotide sequences can lead to a better resolution of the tree. The DNA sequences are used for phylogenetic analyses of closely related species because of more information

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

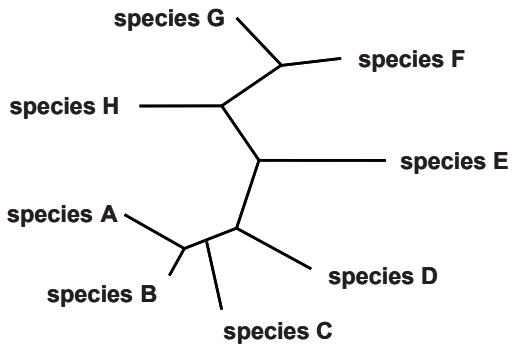

Figure 1. An example of unrooted tree

at the DNA level when compared with the protein level. Moreover, by coding amino acid characters the potentially informative silent substitutions are ignored (Simmons and Freudenstein 2002). In the case of protein level, there are more possible character states for amino acids as opposed to nucleotides (20 versus 4). The amino acid sequences are used for phylogenetic analyses of more distant relationhips (Simmons and Freudenstein 2002). Generally, they are analysed when transitions and/or third-codon positions are determined to be saturated as indicated by high divergence values between sequences (reviewed in Simmons 2000). However, in some cases the DNA sequences are still used for phylogenetic analyses of distant relationships but it is important to remove the third codon positions as these could present pure noise (reviewed in Baldauf 2003).

The primary source of data used for molecular phylogenetic analyses are sequenced PCR or RT-PCR products, which are either directly used in analyses or translated to amino acid sequences. The obtained PCR (RT-PCR) products can be directly sequenced or it is possible to clone them into a suitable vector, and then to sequence the inserted PCR product. The direct sequencing enables to find some polymorphisms in the sequence, in contrast to the sequencing of cloned PCR product insert

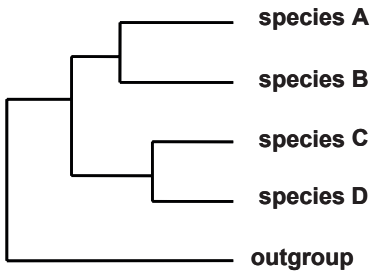

Figure 2. An example of rooted tree

(sequenced plasmids coming from one colony contains only one variant of the gene). Subsequent segregation analysis of the found polymorphism enables to identify whether different alleles or different copies of the gene were found.

If the phylogenetic tree is constructed according to DNA sequences, there is a possibility to choose either repetitive sequences (for example rDNA spacers) or single copy genes for further analyses. The main disadvantage of repetitive sequences is that not all of the repetitive sequences are identical. For example, Desfeux and Lejeune (1996) sequenced rDNA spacer and obtained two types of sequences of *Silene dioica* with completely different branching patterns. Therefore, it is better to sequence more monomers. Furthermore, repetitive sequences between closely related species do not always offer sufficient distinction. On the other hand, phylogenetic analysis of introns of nuclear genes can provide better resolution than repetitive sequences. However, in the case of less conserved sequences, it can be difficult to find orthologs from all studied species by PCR with the same pair of primers.

In general, the genes used for the analysis can have more copies in a respective genome (e.i. they are paralogous), which may cause problems in the construction of phylogenetic trees. For example, it is possible to consider a gene with two copies in all analysed species. If both copies of this gene are detected, the found phylogenetic tree will agree with the true relationships between the species (Figure 4). However, when different copies are found in different species, the resulting phylogenetic tree will not provide the correct relationships between the species (Figure 3) (reviewed in Baldauf 2003). To distinguish between a single copy and a multicopy gene, Southern or PCR analysis can be performed.

In some cases, sequencing of a single gene does not provide the best resolution in the phylogenetic tree. For this reason, it is better to sequence more genes (Sandersson and Driskell 2003). The construction of the phylogenetic tree on the basis of a higher number of sequences, coming from multiple genes, provides generally much more information than a tree constructed according to the sequences coming from one gene.

The second step in the phylogeny reconstruction is to check the sequences from the studied species and to align them. The resulting sequences can be visualized for example using the program BioEdit, available at: http://www.mbio.ncsu.edu/BioEdit/bioedit.html (Hall 1999). Except the pro-

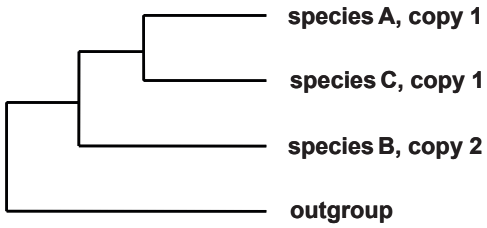

Figure 3. Phylogenetic tree constructed on the basis of different copies of a gene. This tree does not provide the correct relationships between the species

gram BioEdit, it is possible to use the program MEGA3 (Kumar et al. 2004), which I consider not very intuitive. The alignment of the sequences can be performed automatically or manually. Automatic alignments may fail to correctly identify conserved regions, whereas manual alignments allow this, but they are much more laborious. Using a computer-based alignment as a guide to manual alignment offers a good compromise. For example, the programs Clustal W1.81, available at: http://www.cf.ac.uk/biosi/research/biosoft/Downloads/clustalw.html (Thompson et al. 1994); T-Coffee, available at: http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi (Notredame et al. 2000); Muscle, available at: http://www.drive5.com/muscle/ (Edgar 2004); Musca, available at: http://cbcsrv.watson.ibm.com/Tmsa.html (Parida et al. 1998); Mafft, available at: http://www.biophys.kyoto-u.ac.jp/~katoh/programs/align/mafft/ (Katoh et al. 2002); ProbCons, available at: http://probcons.stanford.edu/ (Do et al. 2005) can be used for automatic alignment of sequences. To check and to manually correct the alignments, the program called SeaView, available at: http://pbil.univ-lyon1.fr/software/seaview.html (Galtier et al. 1996) can be used.

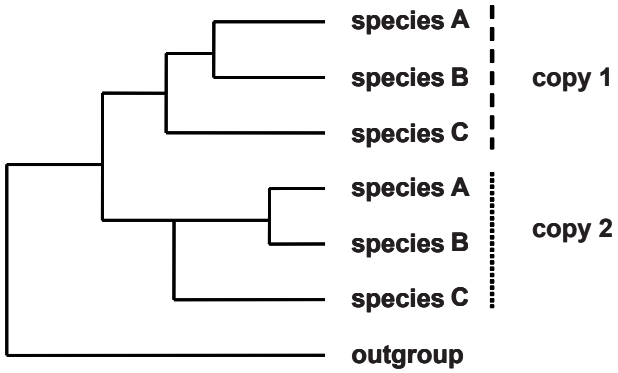

Figure 4. Phylogenetic tree constructed according to all found paralogs of a gene. This tree indicates the correct relationships between the species

Once the data are aligned, there are many different types of phylogenetic analyses, which can be performed (Holder and Lewis 2003). The methods for calculating phylogenetic trees can be generally divided into two categories. These are distance-matrix based methods, also known as clustering or algorithmic methods (e.g. neighbor-joining, Fitch-Margoliash, UPGMA), and discrete data based methods, also known as tree searching methods (maximum parsimony, maximum likelihood and Bayesian methods).

Distance is relatively simple and direct. The distance (roughly, the percent sequence difference), is calculated for all pairwise combinations of operation taxonomic units, and then the distances are gathered into a tree. Discrete data methods examine each column of the alignment separately and look for the tree that best complies all of this information. Unsurprisingly, distance methods are much faster than discrete data methods. However, a distance analysis yields little information other than the tree, while discrete data analyses are information rich. There is a hypothesis for every column in the alignment, so it is possible to trace the evolution at specific sites in the molecule (e.g. catalytic sites or regulatory regions; reviewed in Baldauf 2003).

The most often used methods for phylogenetic analyses are neighbor-joining (NJ), maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference.

NJ is a fast method suited for large datasets. It permits different branch lengths indicating the evolutionary time or amount of evolutionary changes along the branch. The disadvantages of this method are that it gives only one possible tree and it is strongly dependent on the used model of evolution (model is a mathematical description of the sequence evolution and it can be complicated to incorporate other biologically important processes like insertions or deletions; Saitou and Nei 1987). Moreover, the algorithm is based on the reduction of sequence information when transforming the data to the distance matrix, which can be also disadvantageous.

Maximum parsimony method is based on shared and derived characters. It does not reduce sequence information to a single number. It works with original data (alignment) and tries to provide the information about the ancestral sequences. The principle of this method is to find a tree with the smallest number of evolutionary changes (based on the theory that the evolution prefers the smallest number of mutations); in comparison with

the distance based methods it is relatively slow. MP does not use all the sequence information because only informative sites are used, and it does not provide information on the branch lengths. An advantage of the maximum parsimony method is that it does not imply a specific model of evolution and provides all equally parsimonious topologies. A specific problem of MP is long-branch attraction. It is a phenomenon in phylogenetic analyses (especially those using maximum parsimony) caused by the fact that rapidly evolving lineages are considered closely related, regardless of their true evolutionary relationships (reviewed in Bergsten 2005). This problem can be minimized by using methods which comprise differential rates of substitution among lineages or by breaking up long branches by adding taxa that are related to those with the long branches (reviewed in Bergsten 2005), e.g. maximum likelihood. The principle of the latter method is to find such tree-topology, which explains the relationships between the sequences with the highest probability. ML method requires a model of evolution. This is an advantage because it makes us aware of the assumptions being made. The disadvantage of the model is that the use of inadequate likelihood models can lead to interpretation in real data sets. To decide which model best fits the data, the likelihood values given by the different models for the data are calculated and compared. The model to choose is the simplest model that gives a likelihood not significantly lower than the likelihood given by a more general model (reviewed in Felsenstein 2004).

Bayesian inference suggests a natural way how to accommodate uncertainty in phylogenies and provides an intuitive measure of support for trees and a practical way to estimate large phylogenies using a statistical approach (Huelsenbeck et al. 2000). Bayesian inference of phylogenetic trees uses Markov Chain Monte Carlo (MCMC) to approximate the posterior probabilities of trees (Huelsenbeck and Ronquist 2001). The most important aspect of MCMC Bayesian inference is its computational efficiency. The method allows the incorporation of complex models of the DNA substitution process, and other aspects of evolution. Although Bayesian analysis using MCMC is an elegant method for solving many problems, it is relatively new and there is a number of unsolved questions, e.g. convergence of the Markov Chain, discrepancy between Bayesian posterior probabilities and nonparametric bootstrap test values (Huelsenbeck et al. 2002).

Phylogenetic analysis can be performed for example using the program PhyloWin, available at: http://pbil.univ-lyon1.fr/software/phylowin.html (Galtier et al. 1996). The direct output of a phylogenetic analysis are user-unfriendly formatted files. To visualise and to edit the tree, many different programs can be used, as for instance, NJ plot (Perrière and Gouy 1996, accessible on http://pbil.univ-lyon1.fr/software/njplot.html).

It is important to know how much the individual branches are supported within the tree. Finally the accuracy of resulting phylogeny can be measured by different methods. The most commonly used method is bootstrapping (reviewed in Felsenstein 1985). This technique determines the phylogenetic accuracy and enables to establish a marginal winner among many, nearly equal, alternatives (possibilities). In this method, numerous subsamples (500–2000) are generated. Each of the subsamples has the same size as the original, which is accomplished by allowing repeated sampling of sites. That is random sampling with replacement, constructing trees from each of the subsample and calculating the frequency with which the branching pattern in each of this random subsample is reproduced (Hillis and Bull 1993). The bootstrap represents the value interpreting the number of cases in which the sequences were classified together.

For example, if the species X is found in every subsample tree, then its bootstrap support is 100%. If it is found in only two-thirds of the subsample tree, its bootstrap support is 67%. Generally, the bootstrap values of 70% and higher indicate reliable groupings. When the bootstrap values all over the tree are low, it can indicate problems with long-branch attraction. Then it is possible to remove these sequences from the dataset and observe whether the bootstrap values are increased.

To present phylogenetic trees, there are several widely accepted rules. Branch lengths are almost always drawn to scale. Bootstrap values should be displayed as percentages and only values of 50% and higher are presented, because of easier understanding and comparison with other trees.

To conclude, phylogenetic analysis is a powerful tool for organization and interpreting of molecular data. With even a very basic understanding of general principles and conventions, it is possible to glean valuable information from a phylogenetic tree – on the origin, evolution and possible function of genes and the proteins they might encode (reviewed in Baldauf 2003). I hope that this short article will help to understand basic problems of phylogenetic analysis.

REFERENCES

Baldauf S.L. (2003): Phylogeny for the faint of heart: a tutorial. Trends Genet., *19*: 345–351.

Bergsten J. (2005): A review of long-branch attraction. Cladistics, *21*: 163–193.

Desfeux C., Lejeune B. (1996): Systematics of Euro-mediterranean silene (*Caryophyllaceae*): Evidence from a phylogenetic analysis using ITS sequences. Life Sci., *319*: 351–358.

Do C.B., Mahabhashyam M.S.P., Brudno M., Batzoglou S. (2005): PROBCONS: Probabilistic consistency-based multiple sequence alignment. Genome Res., *15*: 330–340.

Doyle J.J., Gaut B.S. (2000): Evolution of genes and taxa: a primer. Plant Mol. Biol., *42*: 1–23.

Edgar R.C. (2004): MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., *32*: 1792–1797.

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

Felsenstein J. (2004): Inferring Phylogenies. Sinauer Associates, Sanderland, Massachusets.

Galtier N., Gouy M., Gautier C. (1996): SeaView and PhyloWin, two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci., *12*: 543–548.

Hall T.A. (1999): Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98NT. Nucleic Acids Symp. Ser. *41*: 95–98.

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

Hillis D.M., Bull J.J. (1993): An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analyses. Syst. Biol., *42*: 182–192.

Holder M., Lewis P.O. (2003): Phylogeny estimation: traditional and Bayesian inference of phylogenetic trees. Bioinformatics, *17*: 754–755.

Huelsenbeck J.P.H., Larget B., Miller R.E., Ronquist F. (2002): Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol., *51*: 673–688.

Huelsenbeck J.P., Rannala B., Masly J.P. (2000): Accomodating phylogenetic uncertainty in evolutionary studies. Science, *288*: 2349–2350.

Huelsenbeck J.P., Ronquist F. (2001): MRBAYES: Bayesian inference of phylogeny. Bioinformatics, *17*: 754–755.

Katoh K., Misawa K., Kuma K., Miyata T. (2002): MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., *30*: 3059–3066.

Kumar S., Tamura K., Nei M. (2004): MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment briefings. Bioinformatics, *5*: 150–163.

Notredame C., Higgins D., Heringa J. (2000): T-Coffee: A novel method for multiple sequence alignments. J. Mol. Biol., *302*: 205–217.

Parida L., Floratos A., Rigoutsos I. (1998): Musca: An algorithm for constrained alignment of multiple data sequences. In: Proc. 9th Workshop on Genome Informatics, Tokyo, Japan.

Perrière G., Gouy M. (1996): WWW-Query: An online retrieval system for biological sequence banks. Biochimie, *78*: 364–369.

Saitou N., Nei M. (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., *4*: 406–25.

Sanderson M.J., Driskell A.C. (2003): The challenge of constructing large phylogenetic trees. Trends Plant Sci., *8*: 374–379.

Simmons M.P. (2000): A fundamental problem with amino-acid-sequence characters for phylogenetic analyses. Cladistics, *16*: 274–282.

Simmons M.P., Freudenstein J.V. (2002): Artifacts of coding amino acids and other composite characters for phylogenetic analysis. Cladistics, *18*: 354–365.

Soltis D.E., Albert V.A., Savolainen V., Hilu K., Qiu Y.L., Chase M.W., Farris J.S., Stefanović S., Rice D.W., Palmer J.D., Soltis P.S. (2004): Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci., *9*: 477–483.

Thompson J.D., Higgins D.G., Gibson T.J. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., *22*: 4673–4680.

Received on February 27, 2007

Corresponding author:

Mgr. Elleni Michu, Akademie věd České republiky, Biofyzikální ústav, v. v. i., Královopolská 135, 612 65 Brno, Česká republika
phone: + 420 541 517 111, fax: + 420 541 240 500, e-mail: michu@ibp.cz