Adv Biochem Eng Biotechnol (2017) 160: 185–196
DOI: 10.1007/10_2016_49

Published online: 26 October 2016

Molecular Phylogenetics: Concepts for a Newcomer

Pravech Ajawatanawong

Abstract Molecular phylogenetics is the study of evolutionary relationships among organisms using molecular sequence data. The aim of this review is to introduce the important terminology and general concepts of tree reconstruction to biologists who lack a strong background in the field of molecular evolution. Some modern phylogenetic programs are easy to use because of their user-friendly interfaces, but understanding the phylogenetic algorithms and substitution models, which are based on advanced statistics, is still important for the analysis and interpretation without a guide. Briefly, there are five general steps in carrying out a phylogenetic analysis: (1) sequence data preparation, (2) sequence alignment, (3) choosing a phylogenetic reconstruction method, (4) identification of the best tree, and (5) evaluating the tree. Concepts in this review enable biologists to grasp the basic ideas behind phylogenetic analysis and also help provide a sound basis for discussions with expert phylogeneticists.

Keywords Evolutionary trees, Molecular phylogenetics, Phylogenetic analysis, Phylogenetic markers, Phylogeny

Contents

P. Ajawatanawong (✉)
Department of Microbiology, Faculty of Science, Mahidol University, 272 Rama VI,
Rachathewi, Bangkok 10400, Thailand
e-mail: pravech.aja@mahidol.ac.th

1 Introduction

Biological research has changed rapidly in the last decade, particularly following the launch of next-generation sequencing (NGS) technology [1]. This is because NGS dramatically reduces sequencing prices, speeds up the process, and generates high-throughput DNA sequencing results. Moreover, the advancements in several "*-omic*" areas also drive biological research in the new era of bioinformatics, systems biology and networking biology. Because the generation of data is easy today, the bioresearch paradigm has shifted from the generation of sequence data to analysis efficacy and power.

Molecular phylogenetics is a disciplinary study of evolutionary relationships amongst organisms using molecular sequences. The analysis methods used in molecular phylogenetics were originally developed to reveal evolutionary pathways, yet today molecular phylogenetics is used in several fields, such as systematic biology and biodiversity [2], molecular epidemiology [3–5], identification of gene functions [6], and microbe identification in microbiome studies [7–9]. For these reasons, molecular phylogenetics is a fundamental field in science of which most biologists require background knowledge.

This review aims to introduce network biologists who are new to the field of molecular phylogenetics to the basic concepts and ideas behind phylogenetic analysis. It begins with the frequently used terminology, characteristics of sequencing markers, and general methods for tree reconstruction and tree evaluation. It then discusses some popular computer programs and critical points that need to be considered in the analysis.

2 Phylogenetic Tree

A phylogenetic tree or phylogeny is a tree-like diagram used to visualize evolutionary relationships among a set of operational taxonomic units (OTUs). The OTU generally represents a species, but can also represent individual organisms in a population, a gene or protein sequence or a taxon at any taxonomic rank (e.g., family, order, class, phylum). The tree is composed of nodes and branches (Fig. 1). Nodes at the tips of the tree are called '*external nodes*.' These are used to represent the OTUs. Another type of node, called '*internal nodes*,' represents a recent common ancestor (RCA). Between these are lines, called '*branches*,' used to connect newer and older nodes and show the evolutionary relationships among

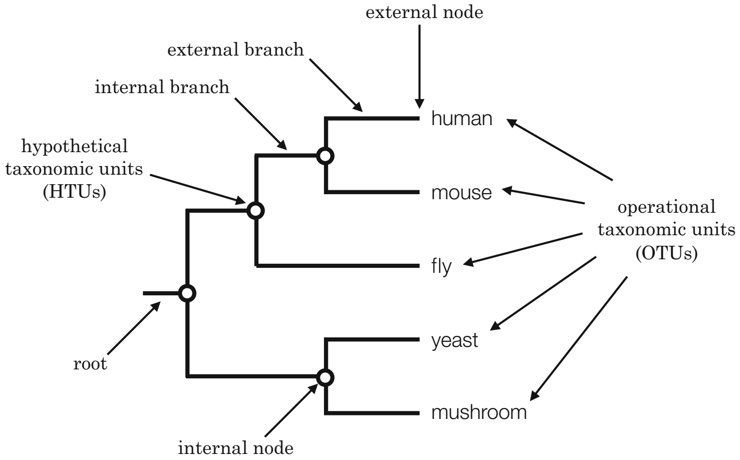

Fig. 1 Composition of a phylogenetic tree. Terminology frequently used in phylogenetic trees is labeled on the tree

the taxa. A branch linking two internal nodes is an '*internal branch*,' which shows an ancient relationship. Conversely, the branch joining an internal node with an external node to show a modern relationship is called an '*external branch*.'

The deepest branch of the tree represents the '*root*' or the '*most recent common ancestor*' (MRCA) of all taxa in the tree. Generally, phylogenetic software can only reconstruct an '*unrooted tree*' or a tree showing who is closely related to whom. To give the tree more meaning in an evolutionary context, the '*rooted tree*' is reconstructed by identifying the origin of all taxa. The best way to root a phylogenetic tree is by adding an '*outgroup*' in the dataset. Theoretically, the root of the tree is located between the outgroup and the remaining taxa. So the best outgroup is an organism or group of organisms recently diverged from the remainder of the organisms in the tree. If an outgroup is unknown or if an ideal outgroup is unavailable (e.g., if there are no data or closely related specimen available), the middle point of the longest branch on the tree can be used as the root of the tree.

The branching pattern dividing the two new nodes is called a '*bifurcation*' or a '*dichotomy*.' This fits with the concept of speciation, in which organisms split from one ancestor into two new species. The tree that contains only bifurcating nodes is a '*fully resolved tree*.' If a deeper node branches into more than two new nodes (three or more), this branching pattern is said to be '*multifurcating*' or a '*polytomy*.'

To read a tree properly, one needs to understand that all branches on the tree can be rotated around a node while retaining the same meaning in the context of evolutionary relatedness (Fig. 2a). Sometimes unrooted phylogenies are drawn in a star-like shape, also called a '*star tree*.' In this case, all branches can be rotated too (Fig. 2b) and the angles of all nodes are meaningless. A phylogenetic tree clusters taxa based on their evolutionary relationships. The closely related taxa are grouped together and share an RCA, whereas more distantly related taxa share a deeper (earlier) common ancestor. All taxa that are descended from the same ancestor

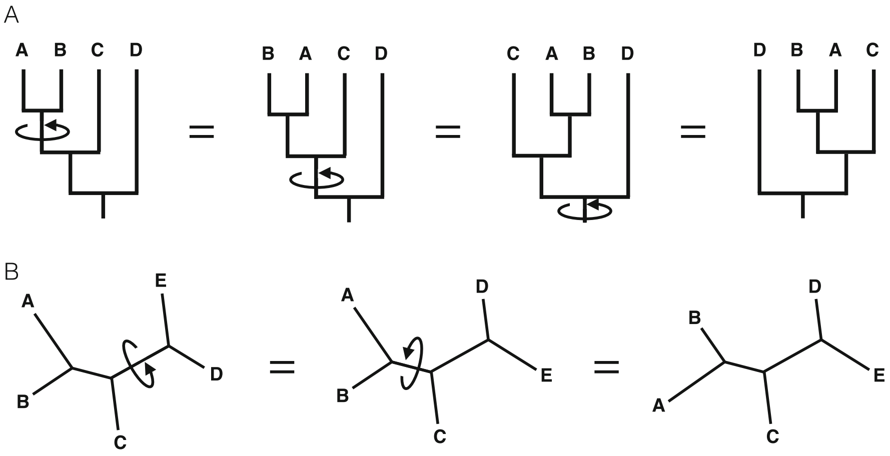

Fig. 2 A phylogenetic tree is similar to a mobile. Rotating the branches on the tree does not change the topology (branching pattern) and meaning of the evolutionary relationship in both rooted (**a**) and unrooted (**b**) phylogenies

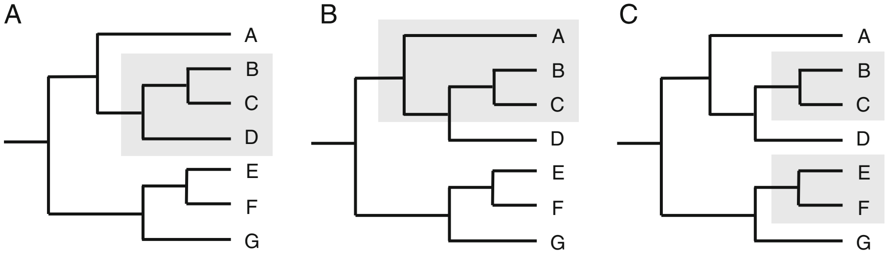

Fig. 3 Examples of a monophyletic group (**a**), a paraphyletic group (**b**), and a polyphyletic group (**c**) are shown in *gray*

make up a '*monophyletic group*' or '*clade*' (Fig. 3a). However, a group of organisms that shares the same ancestor, but does not include all members descending from that ancestor, is called a '*paraphyletic group*' or '*glade*' (Fig. 3b). Another type of group in phylogenetics is a '*polyphyletic group*' (Fig. 3c). This term refers to a group of taxa that are homoplasy. This means that they are not derived from the same ancestor and the term is usually uses for describing convergent evolution.

Molecular phylogenetic analysis must begin with a set of homologous sequences. The homology in molecular sequences is based on the sequences being derived from the same ancestor. With this in mind, molecular homology can be classified into three different types based on genetic mechanisms that separate the daughter sequences. The first type is '*orthologous genes*.' This means that the sequence was once present in the genome of an ancestor, and was transferred to the new species by speciation. This kind of gene is potentially informative for molecular phylogeny. Conversely, some genes are duplicates of other genes in the same genome and are called '*paralogous genes*.' They can cause confusion in the tree reconstruction. Finally, the type of homologous genes that must be avoided in molecular phylogenetic reconstructions are '*xenologous genes*.' These arise from horizontal gene transfer from one species to another. This type of gene can be problematic for a gene tree reconstruction and so usually are best avoided.

3 Molecular Markers for Building a Tree

Over the last few decades, DNA sequences have been accepted and widely used as molecular characters for phylogenetic tree reconstructions, surpassing the use of morphological characters [10]. This is because the sequence states of DNA, which can be only adenine, thymine, cytosine, or guanine, are clearer than morphological states. Molecular sequences also provide a large number of characters for phylogenetic analysis. For example, a phenotype regulated by single gene or a group of genes can be recognized as one character, but almost all positions in a gene's DNA sequence are useful characters for phylogenetic analysis. In addition, sequence-based phylogeny allows scientists to compare organisms across higher taxonomic ranks, such as class, phylum, or even kingdom, despite a lack of comparable morphology (see [11] for further discussion).

Ribosomal RNA (rRNA) sequences in the small subunit (SSU) of the ribosome (16S rDNA sequences for prokaryotes and 18S rDNA sequences for eukaryotes) are the most widely used molecular region for phylogenetic analyses [12–15]. There are several reasons why the SSU is a very powerful marker [16, 17]. First, it is an ancient molecule which emerged during the very early stages of life and it codes for a function necessary for the survival of all cellular organisms. It is therefore present in all organisms. This allows different organisms with no morphology in common to be compared. Second, this molecule is vertically transferred with a low rate of mutation. This means the SSU is very conserved in its sequence, structure, and function. Third, the SSU sequence has multiple variable regions (V1–V9) which are all flanked with conserved blocks. This is convenient for finding oligonucleotide primers to amplify a piece of the SSU DNA for testing the diversity of sequences. In addition to the SSU, there are many other sequence markers which are potentially useful and have been used for phylogenetic analyses. Generally, a potentially useful marker sequences should be single copy and located in either the genome of the nucleus or organelles [18] such as the mitochondrial or plastid genomes (see further details in [19]). They can be either coding or non-coding sequences.

The tree built from a gene is called a '*gene tree*.' Normally, a gene tree can illustrate the evolutionary history of that gene, which is not necessarily the same as the story of the species' evolution. As such, it is probable that the topology (branching pattern) of the gene tree might not be identical to the '*species tree*.'

Phylogenetic tree reconstruction based on multiple genes is an alternative way to improve the resolution of a gene tree and avoid the biases that come with a tree generated from a single gene. Phylogenetic signals from different genes can be combined by concatenating all the aligned sequences. This approach aims to integrate the signal from each gene to make it more intense.

Rokas and Holland [20] proposed the term '*rare genomics changes*' (RGCs), which refers to regions in the genomes of organisms in a particular clade that have rare mutational changes, which can be used as novel markers in molecular phylogeny and evolution. Some examples of RGCs include indels (insertions/deletions), sequence signatures, and amino acid composition changes, which show the potential of RGCs as evolutionarily informative markers [21–24]. There have been some attempts to use RGCs as data for phylogenetic tree reconstruction, but it is very difficult to measure the rate of evolution in these markers and there is also no accepted weighting method for them.

4 Sequence Alignment

DNA and protein sequences are the most frequently used data types in molecular phylogenetic analysis. To study deep phylogeny, one needs ancient, universal, orthologous sequences to form the dataset. However, these sequences might be very diverse and may not align properly. To circumvent this problem, protein sequences are a better choice. This is because mutations appear to have fewer effects on protein sequences. On the other hand, the study of recent evolution or phylogenetic analysis of OTUs within the same species needs DNA sequences, which are less conserved in their sequences than are proteins. Moreover, the analysis of non-coding sequences can be carried out on DNA sequences only.

Molecular phylogenetic analysis relies heavily on the accuracy of the sequence alignment. The programs used for the alignment of sequences are developed from several algorithmic approaches. One of the most popular algorithms is '*progressive sequence alignment*,' which has been implemented in several software packages, such as MUSCLE [25, 26], MAFFT [27, 28], and Clustal Omega [29]. The general concept on which progressive sequence alignment is based is the construction of a '*guide tree*,' which is not meant to be accurate. The guide tree is used to identify sequences with the highest similarity to align first. That is because they are the easiest sequences to align. Then the algorithm keeps adding less similar sequences to the previous alignment. If a gap is needed, it is inserted into the previous sequence alignment and added to all sequences. Once all sequences are aligned, a better tree, which is built from more sophisticated methods, is created and used as a guideline for improving the final alignment.

Most alignment algorithms were developed to perform a good alignment of conserved regions, but none are powerful enough to handle indel (insertion/deletion) regions properly. Moreover, most tree reconstruction methods are developed based on substitution models. Therefore, all indel regions should be removed from the alignment to avoid errors in the analysis. There are some programs that can identify conserved regions and help the user eliminate indel regions from an analysis, such as SeqFIRE [30] and GBLOCKS [31, 32].

5 Phylogenetic Reconstruction

Methods for phylogenetic reconstruction can be classified into two main approaches: distance-based methods and character-based methods. The concept behind the former is the transformation of all sequence information into a distance matrix, which is then analyzed using an algorithm for clustering the taxa. Building a tree with this method is fast but all sequence information is lost in the process. The latter method is time-consuming because all the sequence information is used for the evaluation of the best phylogenetic tree. The calculation of phylogenetic trees using this method can be carried out using several approaches, such as maximum parsimony (MP), maximum likelihood (ML), or Bayesian analyses.

5.1 Distance-Based Approach

The key concept behind distance matrix methods is the conversion of a pairwise sequence alignment into distant values. Because a multiple sequence alignment (MSA) must contain three or more sequences, distance values from all possible pairwise sequences generate a distance matrix. Once a matrix is developed, the alignment is no longer used for the phylogenetic reconstruction. At this point, the matrix is used as the input for the tree building. Different tree building approaches used include the unweighted pair group method with arithmetic mean (UPGMA), weighted pair group method with arithmetic mean (WPGMA), neighbor-joining (NJ), least square (LS), and minimum evolution (ME) methods.

To infer sequence evolution, substitution models are used to calculate a distance value. The simplest method, which can infer the distance from both nucleotide and protein sequences, is *p*-distance. This is based on the level of sequence similarity for each pair in the alignment. Jukes–Cantor's one-parameter (JC69) model assumes that all changes in nucleotides occur at the same rate [33], whereas Kimura's two parameters (K80) model treats the occurrence of transitions and transversions as different rates [34]. The JC69 and K80 models both assume nucleotide substitution moves toward an equilibrium, which means the frequency of each nucleotide is close to 0.25. In the case of disequilibrium, one needs to employ another substitution model which fits the observed mutations. Some other models include F81 [35], HKY85 [36], TN93 [37], and more (see details in [38, 39]). Using an appropriate model for phylogenetic tree reconstruction is important to avoid errors in the clustering step. There are a number of software packages used for testing the applicability of the relevant model against the MSA, such as ModelTest [40] and jModelTest [41].

It is more complicated to infer protein substitutions. This is because changes in protein sequences result from substitutions in the DNA. However, there have been some attempts to observe amino acid substitutions in protein sequences by using a protein substitution matrix. There are two main matrix approaches generally used in sequence analysis software, including those used in phylogenetic analysis. One of these is called the percentage accepted mutation (PAM) matrix [42] and the other is the blocks substitution matrix or BLOSUM [43]. The PAM models with a higher number (e.g., PAM250) and the lower number BLOSUM matrices (e.g., BLOSUM30) are suitable for more diverse amino acid sequences, whereas the PAM models with a lower number (e.g., PAM60) and the higher number BLOSUM matrices (e.g., BLOSUM90) are suitable for the highly conserved amino acid sequences.

The major advantage of distance matrix methods is their rapid calculation speed. This is possible because the method dramatically reduces the amount of data from a long sequence alignment into a single distance matrix. Moreover, this method may give reliable results if homoplasy is rare and randomly distributed throughout the tree. However, reduction of the data leads to a loss of sequence information and can sometime generate negative branch lengths, which lack biological meaning. Instead, distance-based approaches (e.g., the NJ method) are recommended for large datasets ($>$1,000 sequences) with high sequence similarity.

5.2 Character-Based Approach

There are several methods that have been developed from character-based approaches, such as maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference methods. These approaches aim to reconstruct a phylogeny directly from the sequence data, without any transformation. They make extremely slow calculations but the final tree is said to be very accurate. Briefly, the algorithm used in these begins with scoring all possible phylogenies that can be generated from the *n* taxa. Then the optimal tree is assumed to be the tree with the best score. However, it is nearly impossible to score all of the individual trees when the number of taxa is larger than 20 (as this means the number of possible trees is larger than 2.21×10^{18}) by using a greedy method that searches all possible trees. Some computational search algorithms allow the user to score and select from all possible trees simultaneously. They also reduce the number of possible trees by skipping the theoretically impossible topologies from the possible trees, resulting in an increased search speed. Two popular search algorithms, which are implemented in most current phylogenetic software, are the '*branch-and-bound*' and '*heuristic*' methods. The process of the former method starts with the generation of a core tree: a three-taxa phylogeny. Then a random new taxon from the dataset is added into the core tree, and the only the new trees with an improved score have the fifth taxon added to them. This process is continued until the algorithm reaches the last taxon. The heuristic method is generally similar to the branch-and-bound method, but instead of adding new taxa into the tree with an improved score over the previous tree, the heuristic method uses only the tree with the best score in each round of taxon addition.

The maximum parsimony (MP) method—the oldest phylogenetic method—is a substitution model-free method for phylogenetic tree reconstruction. It is mostly used for building trees from morphology-based data, where it is difficult to measure the rate of evolutionary change. When this method is applied to molecular sequences, each column in the MSA is treated as a individual character. Even though each molecular sequence contains numerous characters, not every position is useful in the MP analysis (e.g., invariable sites). Characters (columns in the MSA) having at least two states (more than two types of nucleotide or amino acid) are called '*parsimony informative sites*,' and only these are included in the MP analysis. The MP method searches for the '*the most parsimonious tree*' or '*the maximum parsimony tree*,' which requires the minimum number of steps to build. Phylogenetic tree reconstruction using this method can give a reliable result if homoplasy occurs in the sequence data either randomly or infrequently. Moreover, this method can be easily applied to any novel type of data, such as indel positions. However, most sequences do not simply evolve at a low rate, and as a result sequences can be difficult to align, which makes MP less efficient, particularly when alignment patterns are complicated. MP is a time-consuming method, and it is not recommended when multiple-gene sequences are concatenated or with sequences with high levels of variation [44].

The second popular method for phylogenetic tree reconstruction is maximum likelihood (ML). ML is a statistical method used to estimate the parameters of a model given the data, and was first applied in phylogenetic analyses of DNA and protein sequences by Felsenstein [35]. In phylogenetic analysis, the ML method estimates the branch lengths and topology of the tree based on the substitution model and the sequence alignment. The numerical output of the ML analysis is the probability that a tree topology and model fit to the sequences. The calculation is repeated for all possible tree topologies that can be generated from n taxa. The tree topology with the highest maximum likelihood value is then reported as the best tree or the '*maximum likelihood tree*.' The strong point of the ML method is that it is claimed to be very accurate. This is because the analysis relies heavily on the evolutionary model. Because of this, all substitution models that can be used in the distance matrix methods can also be used for tree selection. Unlike the MP method, the ML method uses all the information in the sequences to calculate the maximum likelihood value. However, this results in a slow calculation time. Likewise, another weak point of the ML method is that it is impractical for large data sets. This is because the calculation is robust and requires significant computational resources.

Bayesian statistics is the newest method, which was first used for phylogenetic tree reconstruction about two decades ago [45]. This method depends on Bayesian statistics, and aims to search for the tree that maximizes the chance of seeing the model given the data (see details in [46–48]). In brief, the Bayesian phylogenetic algorithm searches for the tree that has the highest posterior probability. To deal with the enormous number of possible trees, Bayesian phylogenetic inference uses a Markov chain Monte Carlo (MCMC) algorithm to search for the best tree. This technique is more sophisticated than that used in the ML method because every new tree that is explored can produce a lower score than the tree in the previous step. This allows the Bayesian inference algorithm to find the best tree efficiently. There are some popular programs that implement the Bayesian inference algorithm, such as MrBayes [49, 50], PhyloBayes [51, 52], and BEAST [53].

Once a tree is reconstructed it is necessary to visualize it. There are no set rules for presenting a tree, but using color and renaming taxa to something easy to understand are always beneficial to the reader. Generally it is best to try to avoid using sequence codes or accession numbers to label OTUs. Likewise, it is critical to write a summary of the method used to build the tree to present in the figure legend. This helps the user to understand the tree more easily [54].

6 Conclusion

Phylogenetic analysis is one of the important techniques in the networking biologist's toolbox. It can be used to identify the evolutionary relationships among organisms, as well as gene or protein sequences. To analyze an evolutionary pathway, one needs to start with orthologous sequences and perform the analysis properly. However, single gene phylogenies generally have less evolutionary signal. As genomes are now being widely sequenced, the possibility of tree reconstruction based on entire or nearly complete genomes is emerging. This approach may replace traditional techniques in molecular evolution in the near future.

Acknowledgement I thank Prof. Prasit Palittapongarnpim and Dr. Nancy Irwins for comments on this review. This work was funded by Mahidol University.

References

1. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141
2. Senés-Guerrero C, Schüßler A (2016) A conserved arbuscular mycorrhizal fungal core-species community colonizes potato roots in the Andes. Fungal Divers 77:317–333
3. Baele G, Suchard MA, Rambaut A, Lemey P (2016) Emerging concepts of data integration in pathogen phylodynamics. Syst Biol. p ii: syw054
4. Bentley SD, Parkhill J (2015) Genomic perspectives on the evolution and spread of bacterial pathogens. Proc Biol Sci 282:20150488
5. Kenah E, Britton T, Halloran ME, Longini IM Jr (2016) Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. PLoS Comput Biol 12, e1004869

6. Chang AB, Lin R, Keith Studley W, Tran CV, Saier MH Jr (2004) Phylogeny as a guide to structure and function of membrane transport proteins. Mol Membr Biol 21:171–181
7. Carrillo-Araujo M, Taş N, Alcántara-Hernández RJ, Gaona O, Schondube JE, Medellín RA, Jansson JK, Falcón LI (2015) Phyllostomid bat microbiome composition is associated to host phylogeny and feeding strategies. Front Microbiol 6:447
8. Martiny JBH, Jones SE, Lennon JT, Martiny AC (2015) Microbiomes in light of traits: a phylogenetic perspective. Science 350:aac9323
9. Matsen FA (2015) Phylogenetics and the human microbiome. Syst Biol 64:e26–e41
10. Lumbsch HT, Leavitt SD (2011) Goodbye morphology? A paradigm shift in the delimitation of species in lichenized fungi. Fungal Divers 50:59–72
11. Hillis DM (1987) Molecular versus morphological approaches to systematics. Ann Rev Ecol Syst 18:23–42
12. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73:278–288
13. Ettoumi B, Guesmi A, Brusetti L, Borin S, Najjari A, Boudabous A, Cherif A (2013) Microdiversity of deep-sea *Bacillales* isolated from Tyrrhenian sea sediments as revealed by ARISA, 16S rRNA gene sequencing and BOX-PCR fingerprinting. Microbes Environ 28: 361–369
14. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173:697–703
15. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12: 635–645
16. Nadler SA (1995) Advantages and disadvantages of molecular phylogenetics: a case study of ascaridoid nematodes. J Nematol 27:423–432
17. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9, e93827
18. Zhang N, Zeng L, Shan H, Ma H (2012) Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol 195:923–937
19. Patwardhan A, Ray S, Roy A (2014) Molecular markers in phylogenetic studies—a review. J Phylogenetics Evol Biol 2:2
20. Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459
21. Ajawatanawong P, Baldauf SL (2013) Evolution of protein indels in plants, animals and fungi. BMC Evol Biol 13:140
22. Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A 90:11558–11562
23. Chernikova D, Motamedi S, Csürös M, Koonin EV, Rogozin IB (2011) A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol Direct 6:26
24. Janečka JE, Miller W, Pringle TH, Wiens F, Zitzmann A, Helgen KM, Springer MS, Murphy WJ (2007) Molecular and genomic data identify the closest living relative of primates. Science 318:792–794
25. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113
26. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
27. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

28. Katoh K, Standley DM (2016) A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32:1933–1942
29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539
30. Ajawatanawong P, Atkinson GC, Watson-Haigh NS, Mackenzie B, Baldauf SL (2012) SeqFIRE: a web application for automated extraction of indel regions and conserved blocks from protein multiple sequence alignments. Nucleic Acids Res 40:W340–W347
31. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552
32. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577
33. Jukes TH, Cantor CR (1969) In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 121–123
34. Kimura MA (1980) Simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol 16:111–120
35. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
36. 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
37. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526
38. Liò P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Res 8: 1233–1244
39. Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36: 445–466
40. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818
41. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256
42. Dayhoff MO, Schwartz R, Orcutt BC (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol 5, supplement 3rd edn. Nat Biomed Res Found. pp 345–358
43. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919
44. Stewart CB (1993) The powers and pitfalls of parsimony. Nature 361:603–607
45. Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311
46. Alfaro ME, Holder MT (2006) The posterior and the prior in Bayesian phylogenetics. Annu Rev Ecol Evol Syst 37:19–42
47. Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284
48. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314
49. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 7:754–755
50. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542
51. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Phylogenetics 25:2286–2288
52. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109
53. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973
54. Baldauf SL (2003) Phylogeny for the faint of heart: a tutorial. Trends Genet 19:345–351