Chapter 10
Phylogenetic Trees: Applications, Construction, and Assessment

Surekha Challa and Nageswara Rao Reddy Neelapu

Contents

10.1 Introduction and Applications of Phylogeny

Study of relationships among individuals or groups of organisms or species or populations is called phylogeny. The relationships among the individuals are estimated or assessed based on the evolutionary signals present in the genetic material of any organism. The evolutionary signals or footprints among these individuals or entities are used to construct the evolutionary history. The evolutionary history based on the evolutionary signals can be modeled or represented in the form of graphical

S. Challa · N. R. R. Neelapu (✉)
Department of Biochemistry and Bioinformatics, Institute of Science, Gandhi Institute of Technology and Management (GITAM) (Deemed to be University), Visakhapatnam, Andhra Pradesh, India

K. R. Hakeem et al. (eds.), *Essentials of Bioinformatics, Volume III*,
https://doi.org/10.1007/978-3-030-19318-8_10

representation or tree, which is known as phylogenetic tree. Phylogenetics is an ever-evolving field that promises to give more insights into understanding biodiversity, evolution, ecology, and genomes. Phylogenetics has several applications like affiliating taxonomy to an organism, studying reproductive biology in lower organisms, assessing the process of cryptic speciation in a species, understanding the history of life, resolving controversial history of life, reconstructing the paths of infection in an epidemiology to understand the evolution of pathogen, classifying proteins or genes into families, and many more.

10.1.1 Affiliating Taxonomy to an Organism

Every living organism which is known or identified till date should be classified and affiliated to a taxonomic group. When the taxonomy of the species identified is not known, it is left as an orphan or classified into a special group. The traditional approach for identification of an organism includes studies based on microscopy, morphology, biochemical tests, physiological tests, fruit bodies, mating behavior experiments, and others. The drawbacks associated with the traditional approach are time consuming and of low to moderate in precision. In these cases, phylogeny can be used to affiliate taxonomy to a taxa or an organism.

Phylogeny has been proposed and widely accepted to affiliate taxonomy for a species. Several reports were there on entomopathogenic fungi (Neelapu et al. 2009), *Echinococcus* (Thompson 2008), catfishes (Teugels 1996), *Borrelia burgdorferi* (Margos et al. 2011), *Trichinella* (Pozio et al. 2009), and many more. This case study provides with details that how phylogeny can be used to affiliate taxonomy for entomopathogenic fungi (Neelapu et al. 2009). When the taxonomy of the species is not known, it is left as an orphan or classified into a special group. The fungi which are not classified into any fungal divisions such as *Ascomycota*, *Zygomycota*, and *Basidiomycota* were classified into a special group known as Deuteromycota. Neelapu et al. (2009) studied phylogeny of mitosporic or asexual or conidiogenous entomopathogenic fungi of Deuteromycota belonging to the genera *Beauveria*, *Nomuraea*, *Metarhizium*, *Paecilomyces*, and *Lecanicillium*. One hundred forty-seven fungal entries covering 94 species related to *Ascomycota*, *Zygomycota*, and *Basidiomycota* were analyzed. The partial amino acid sequences of the β-tubulin gene were aligned using AlnExplorer of MEGA ver. 3.014. The statistical procedures minimum evolution (ME), maximum parsimony (MP), and neighbor joining (NJ) of MEGA ver. 3.014; maximum likelihood of PAUP ver. 4b; Bayesian inference of MrBayes ver. 3.04b10; and Metropolis-coupled Markov chain Monte Carlo (MCMCMC) were used to construct phylogenetic tree. "Phylogenetic analysis placed all the asexual entomopathogenic fungal species analyzed in the family *Clavicipitaceae* of the order *Hypocreales* of *Ascomycota*" (Fig. 10.1). Thus, whenever the identity of the organism is in crisis, phylogeny can be used to affiliate the organism to the known traditional taxonomic group.

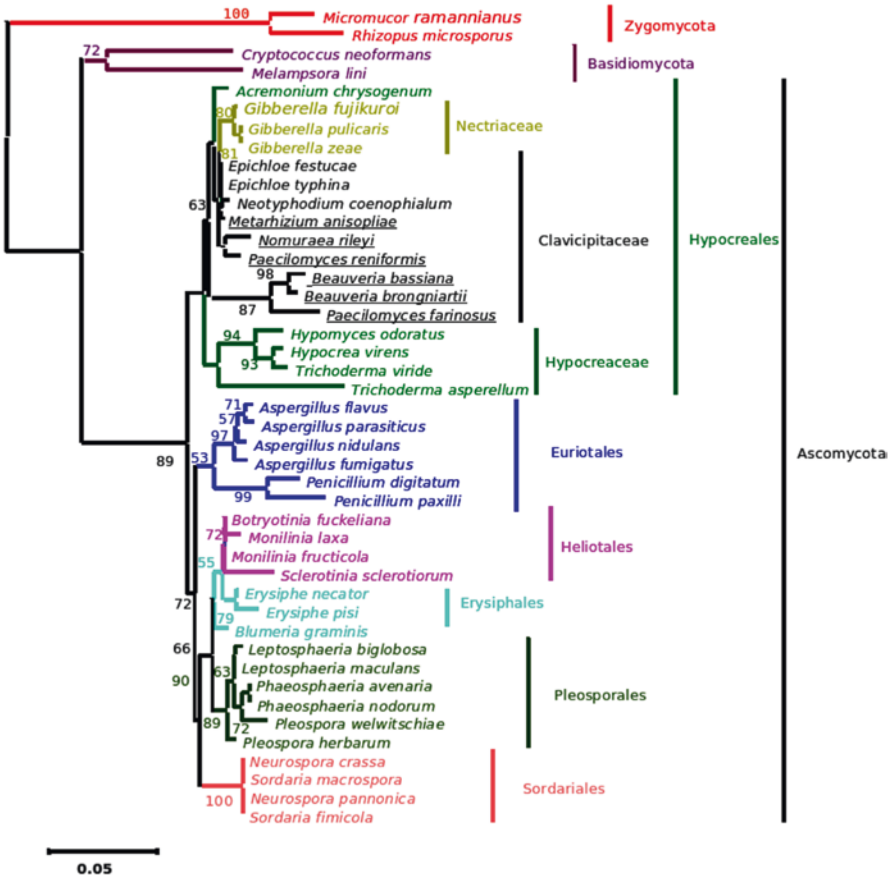


Fig. 10.1 The phylogenetic affiliation of the asexual entomopathogenic *Beauveria* spp., *Nomuraea* spp., *Metarhizium* spp., and *Paecilomyces* spp. (Source: Neelapu et al. 2009)

10.1.2 Studying Reproductive Biology in Lower Organisms

Understanding the reproductive biology in lower organisms where sexual organs are not observed is a challenge. Genetic tests based on phylogenetic concordance and gene genealogies offer an indirect means of identifying recombination. When phylogeny is applied, different genes show different genealogies within a species due to recombination. Therefore, phylogenetic trees generated from the data show phylogenetic concordance among the multiple gene genealogies in recombining species, whereas non-phylogenetic concordance among the multiple gene genealogies in a clonal species (Fig. 10.2).

The reproductive biology in *Beauveria bassiana* (Neelapu 2007; Devi et al. 2006) and *Nomuraea rileyi* (Neelapu 2007; Devi et al. 2007) was studied. Devi et al. (2006)

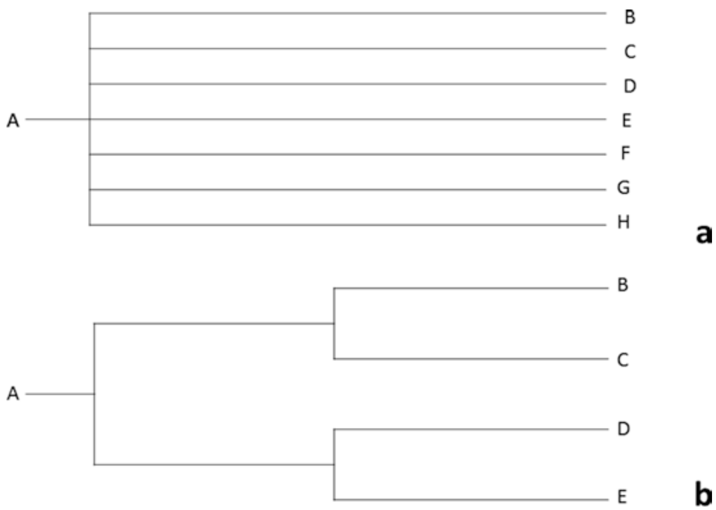


Fig. 10.2 Phylogenetic concordance and gene genealogies: (**a**) clonal species (**b**) recombining species

applied indirect means of genetic tests which are based on phylogenetic concordance of gene genealogies to identify reproductive biology (recombination or clonal) in a localized epizootic population of entomopathogenic fungi *B. bassiana*. Nucleotide sequence data of different allelic forms of three genes (large and small subunits of mitochondrial ribosomal RNA (mt rRNA) and β-tubulin) were evaluated to assess phylogenetic concordance among the multiple gene genealogies. Lack of phylogenetic concordance among three gene genealogies in the epizootic of *B. bassiana* indicates prevalence of recombination within the clonal structure of the population (Fig. 10.3). Thus, whenever the mating tests cannot be applied in lower organisms like bacteria and fungi where sexual organs are not observed, phylogenetic concordance among multiple gene genealogies can be used for understanding the reproductive biology.

10.1.3 Assessing the Process of Cryptic Speciation in a Species

Entomopathogenic fungi of Deuteromycota belonging to the genera *Beauveria*, *Nomuraea*, *Metarhizium*, and *Paecilomyces* are recognized as a "species complex" comprising of genetically diverse lineages. Devi et al. (2006) used amplified fragment length polymorphism (AFLP) and single-stranded confirmation polymorphism (SSCP) data of worldwide population and generated unweighted pair group method with arithmetic mean (UPGMA) tree. The worldwide sample of *B. bassiana* isolates represented cryptic phylogenetic species (Fig. 10.4). Literature reports the use of powerful approach—genealogical concordance phylogenetic species recognition (GCPSR)—to uncover cryptic speciation. "GCPSR detects genetically isolated

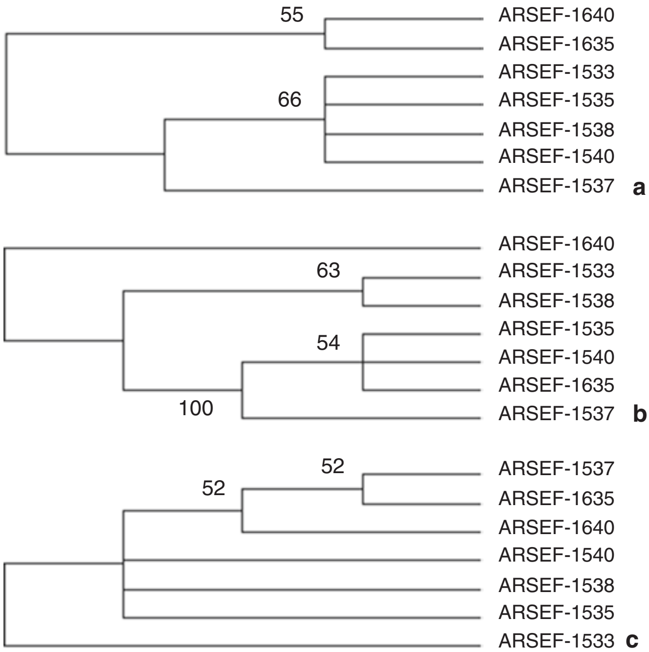


Fig. 10.3 Maximum parsimony tree generated from the sequences of (**a**) partial sequence of β-tubulin gene, (**b**) large subunit of mt rRNA gene, and (**c**) small subunit of mt rRNA genes derived from the isolates of an epizootic *B. bassiana* population from Burgenland, Austria. The tree topology of each species tree indicates the presence of recombination and cryptic speciation. (Source: Devi et al. 2006)

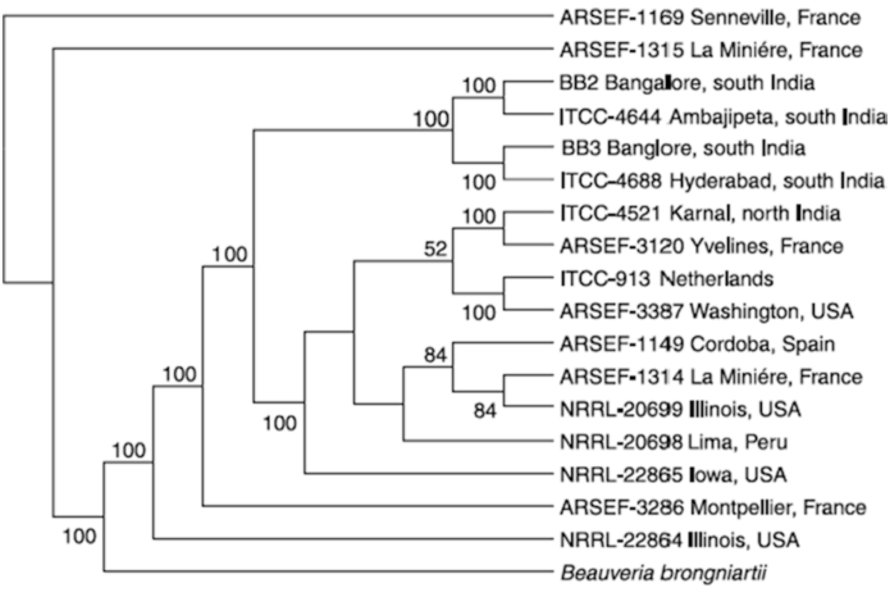


Fig. 10.4 Phylogenetic tree derived and generated from SSCP data of three genes: β-tubulin gene, and large and small subunits of mt rRNA genes of a sample of isolates of *B. bassiana* of worldwide distribution, representing cryptic phylogenetic species (Neelapu et al. 2009)

groups from a number of different loci by comparing the gene trees. Different genes have different genealogies within a species establishing gene flow delimiting species by identifying the unshared polymorphisms, and thus branches that are incompatible, with all genealogies at all loci. Thus, branches that are incompatible with all genealogies at all loci represent different species" (Neelapu et al. 2009).

Neelapu et al. (2009) used GCPSR to uncover cryptic speciation in *B. bassiana*. Epizootic population of *B. bassiana* from Burgenland, Austria, are sequenced for partial sequences of the three genes, β-tubulin gene and large and small subunit of rRNA genes of mitochondria, and were aligned using AlnExplorer of MEGA ver. 3.1. A consensus maximum parsimony tree was generated using PAUP ver. 4.0. "The tree topology of each species tree indicates the presence of cryptic speciation. Incongruity of gene genealogies within a given group indicates gene flow and delimits a species. As the approach detects reproductive isolation, the resulting groups also fulfill the criteria of a biological species" (Fig. 10.3).

10.1.4 Studying the Evolution of Proteins or Gene Families

Phylogeny is used in establishing the origin and evolutionary pattern of a gene of particular species with respect to the other species. Similar set of genes are required for studying or understanding the phylogeny. The genes, which are similar in their structure or function, are known as homologous sequences. If the genes are similar in function but are from different organisms, then they are believed to be orthologous sequences. If the genes are from the same organism, then they are known as paralogous sequences. It is believed that orthologous sequences are due to speciation from a common ancestor, whereas paralogous sequences are due to duplication.

Though there are many reports on the evolution of proteins or gene families, we would like to throw some light on evolution of globin and V-PPases (Hardison 2012; Suneetha et al. 2016). Globin genes diverged to form hemoglobin (oxygen transport in blood), myoglobin (oxygen metabolism in muscle), cytoglobin (oxygen donator during synthesis and cross-linking of collagen or acting as a protector of the free radicals formed in the fibrosis process), and neuroglobin (acts as an oxygen reservoir releasing oxygen in stressful situations, such as hypoxia). So, the plausible explanation for gene evolution can be duplication of the existing gene like globin followed by divergence in function as described above for hemoglobin, myoglobin, cytoglobin, and neuroglobin (Figs. 10.3 and 10.4) (Hardison 2012). The best example for both orthologous and paralogous sequence is globin genes. α-Globin and β-globin genes found in different species are orthologous genes (Fig. 10.5), whereas the α, β, γ, and δ globin genes due to duplication in the same organism are paralogous genes (Fig. 10.6) (Hardison 2012; Opazo et al. 2008).

V-PPase is a heat-stable single polypeptide, coexisting along with V-ATPase on the plant vacuolar membrane in plants, algae, photosynthetic bacteria, protozoa, and archaebacteria (Rea et al. 1992; Maeshima 2000). V-PPase uses ATP and inorganic pyrophosphate (PPi), respectively, as energy sources for generating an electrochemical gradient of protons across the tonoplast. This facilitates the functioning of the Na^+/H^+

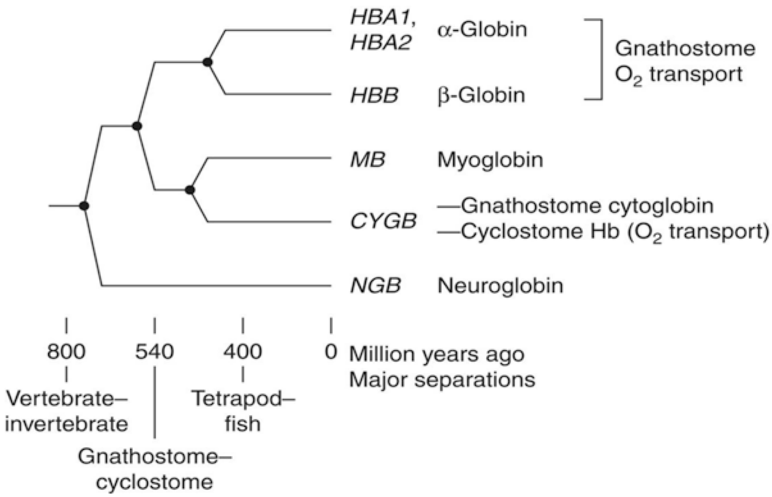


Fig. 10.5 Phylogenetic tree showing duplication and divergence of globin genes, an example for evolution of vertebrate globin genes. (Source: Hardison 2012)

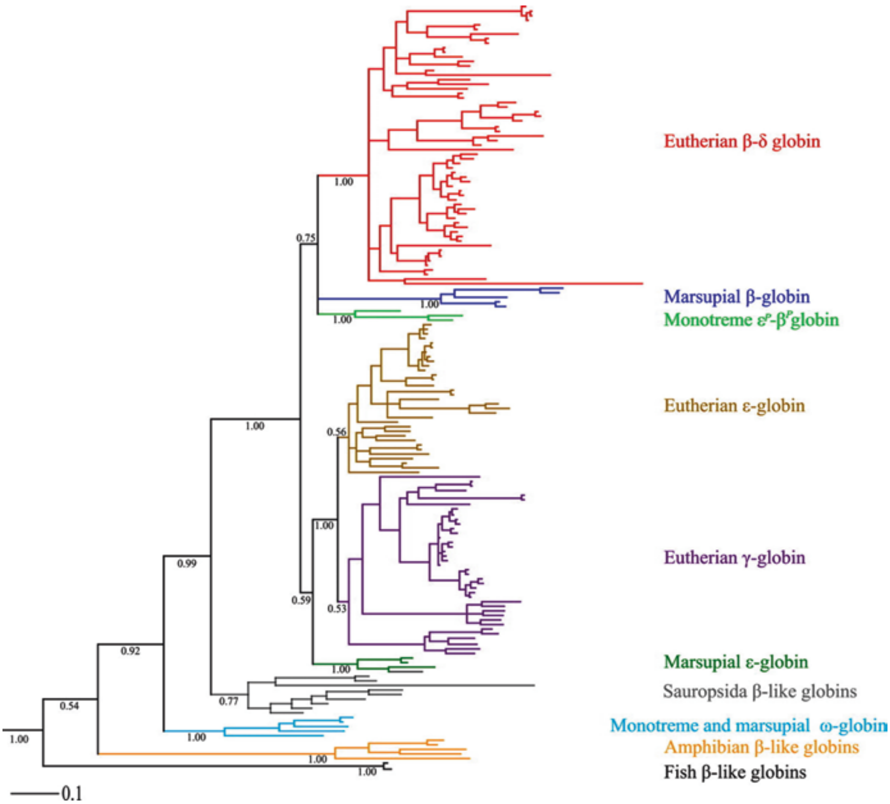


Fig. 10.6 Phylogenetic tree showing relationships among the *β*-like globin genes of vertebrates. (Source: Opazo et al. 2008)

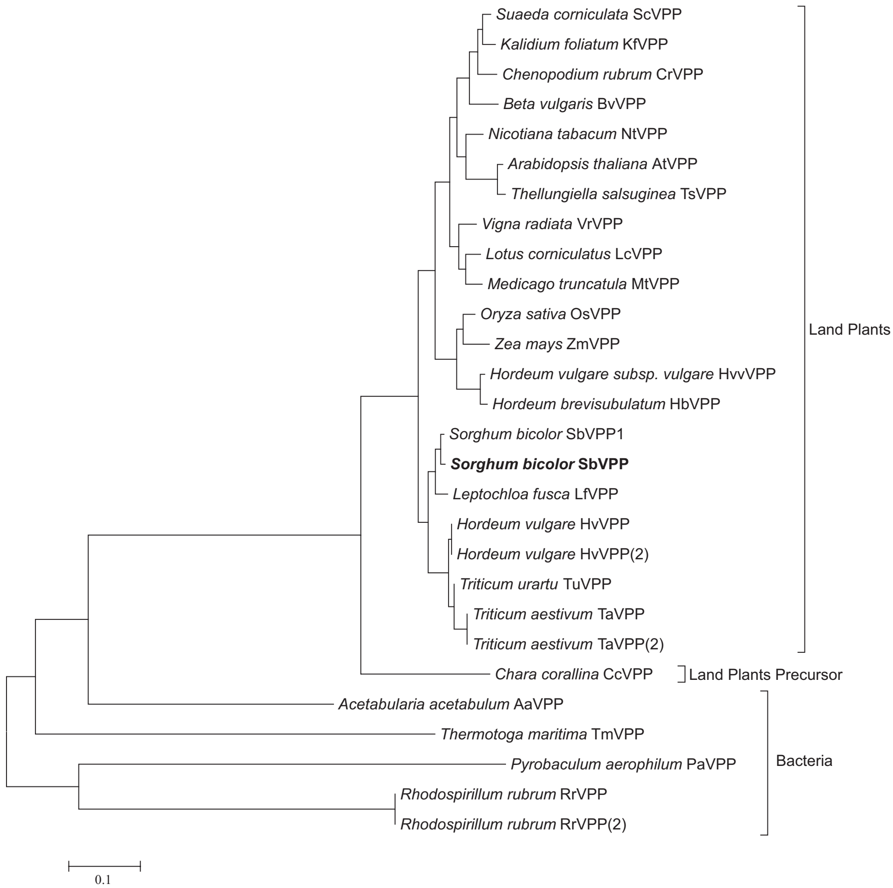


Fig. 10.7 Phylogenetic tree showing relationships among land plants, archae, and bacterial V-PPases. (Source: Suneetha et al. 2016)

antiporter and helps in Na^+ compartmentation. Suneetha et al. (2016) carried out phylogenetic studies on land plants, archaea, and bacterial V-PPases (Fig. 10.7). V-PPases are highly conserved among land plants and less among archaeon, protozoan, and bacteria (Suneetha et al. 2016). Phylogeny with respect to other land plants revealed that V-PPases of *A. thaliana* (AtVPP), *H. vulgare* (HvVPP), *B. vulgaris* (BvVPP), *N. tabacum* (NtVPP), and *O. sativa* (OsVPP) are highly conserved.

10.1.5 Classifying Proteins or Genes into Families

Classification of genes into gene families is important for understanding function and evolution of gene. There are three methods to infer gene families: (1) using phylogenetic trees for classification, (2) using similarities with known sequence signatures like motifs or domains, and (3) pairwise comparisons involving the use of clustering techniques (Frech and Chen 2010).

Phylogenetic tree was used for effective classification of ABC transporter gene families. Multiple sequence alignment of both known and putative new ABC transporter family C genes using ClustalW with default parameters was performed. The phylogenetic tree was produced by the minimum evolution method and 1000 bootstrap iteration. In phylogenetic analysis, the three new genes grouped nicely within known ABC transporters of family C (Fig. 10.8). Thus, phylogenetic analysis can be used to classify new genes into ABC transporter family C (Frech and Chen 2010).

10.1.6 Understanding the History of Life

Understanding the systematics of living organisms in the world is a challenging task. Literature reports several studies carried out to understand the kingdom-level phylogeny. Carl Woese established a molecular sequence-based phylogenetic tree by comparing ribosomal RNA (rRNA) sequences that could relate all organisms and reconstruct the history of life (Woese 1987; Woese and Fox 1997). Woese articulated and recognized three primary lines of evolutionary descent, termed "urkingdoms" or "domains": ....."Eucarya (eukaryotes), Bacteria (eubacteria), and Archaea (archaebacteria)"..... (Woese et al. 1990). Pace (1997) used molecular phylogeny to compile the robust map of life domains: Archaea, Bacteria, and Eucarya (Fig. 10.9). The universal phylogenetic tree based on 64 SSU rRNA sequences was aligned, and a tree was produced using FASTDNAML. Baldauf et al. (2000) used concatenated amino acid sequences of four protein-encoding genes to produce a phylogenetic tree for 14 higher-order eukaryote taxa (Fig. 10.10). Thus, phylogeny was used to understand the kingdom-level relations.

10.1.7 Estimating the Time of Divergence Using Molecular Clock

Molecular dating techniques were used to estimate the time of species divergences. Literature reports several research studies used to determine the time of species divergences. Molecular dating requires standard sequence datasets; statistical distributions to model; and prior divergence times to find out the time of divergence during the course of evolution. Hasegawa et al. (1985) developed a method for estimating divergence dates of humans from species by a molecular clock approach. The molecular clock of mitochondrial DNA (mt DNA) was calibrated ~65 million years ago and a generalized least squares method was applied. The divergence dates were 92.3 ± 11.7, 13.3 ± 1.5, 10.9 ± 1.2, 3.7 ± 0.6, and 2.7 ± 0.6 million years ago for mouse, gibbon, orangutan, gorilla, and chimpanzee, respectively (Figs. 10.11 and 10.12). Thus, phylogeny can be used to estimate time of divergence for species of interest.

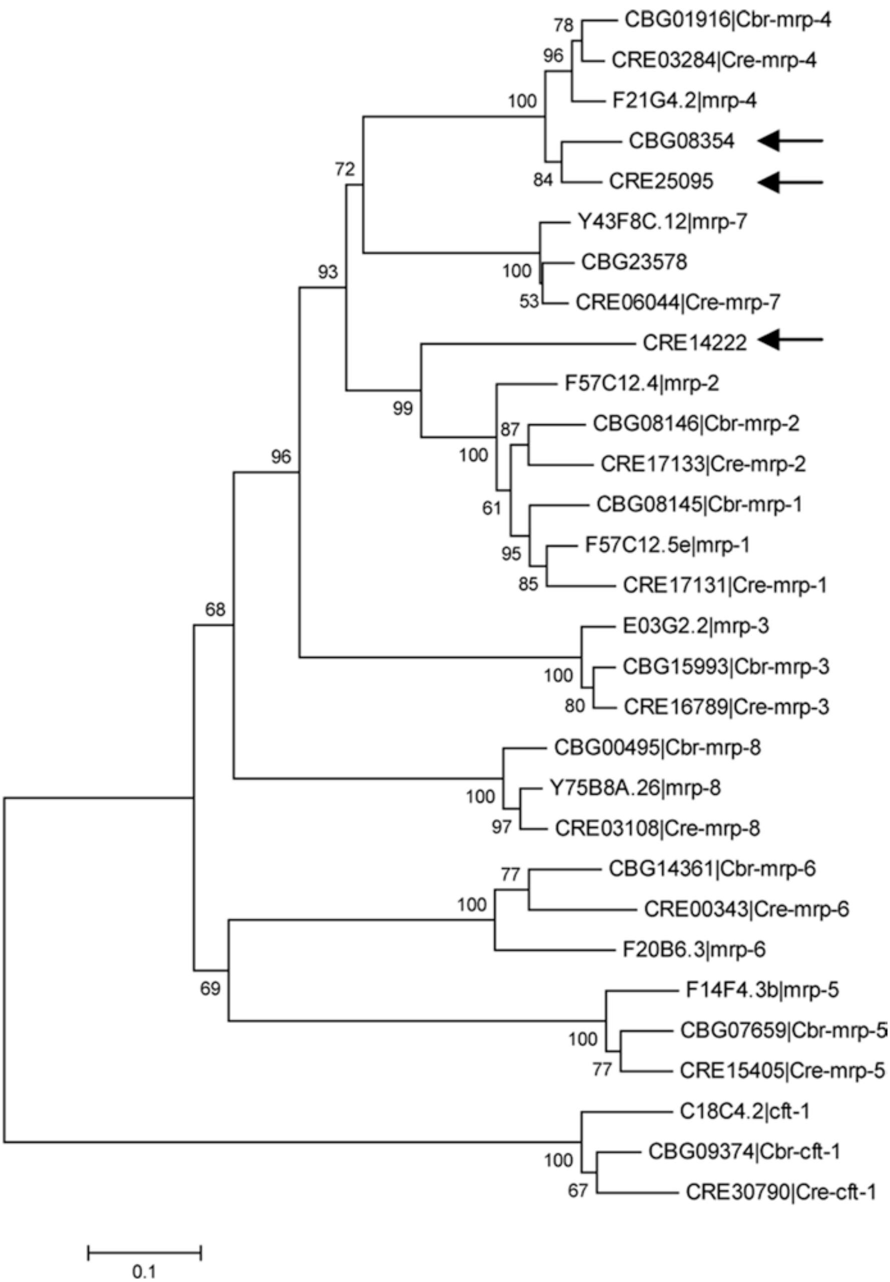


Fig. 10.8 The phylogenetic tree shows the evolutionary relationship of the three new ABC transporter genes CBG08354, CRE25095, and CRE14222 (indicated by arrows) with known *C. elegans*, *C. briggsae*, and *C. remanei* ABC transporters of family C. (Source: Frech and Chen 2010)

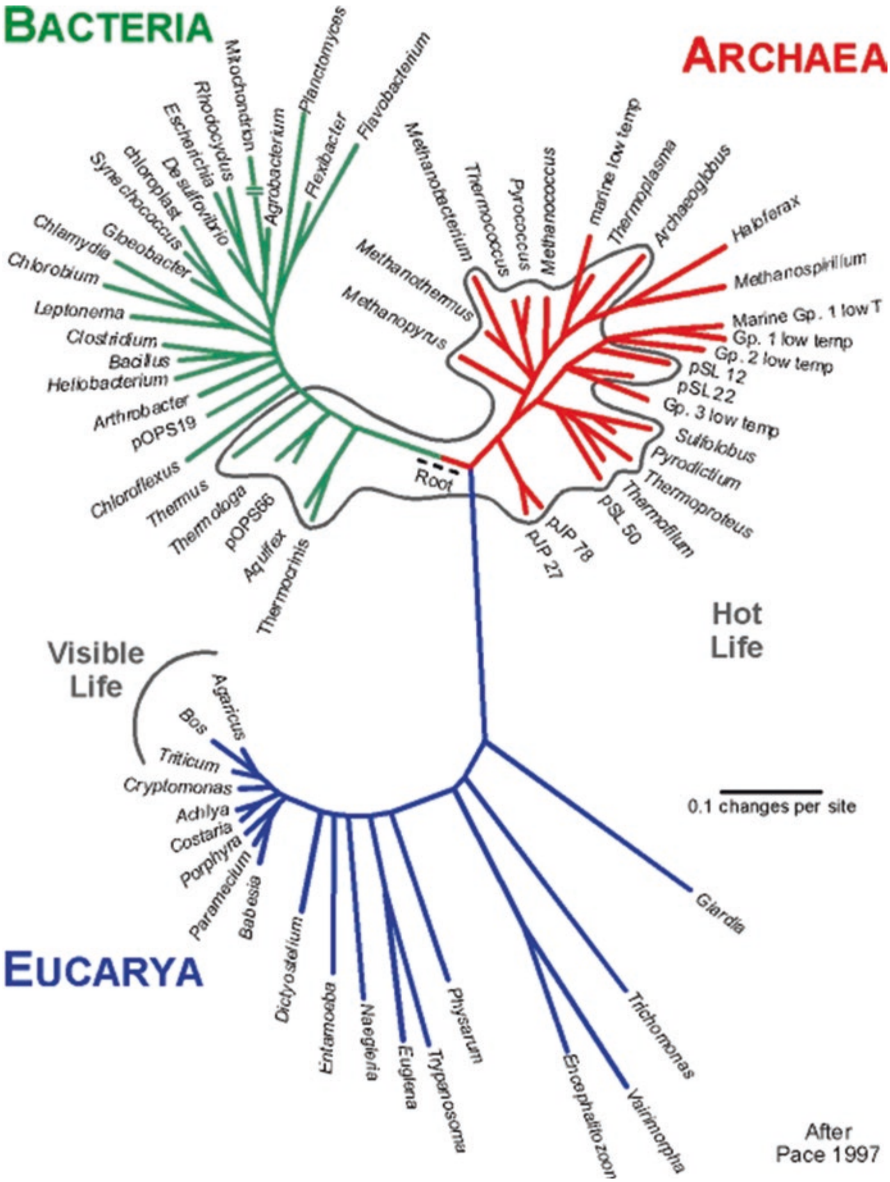


Fig. 10.9 The phylogenetic tree shows the robust map of life domains: Archaea, Bacteria, and Eucarya. (Source: Pace 1997)

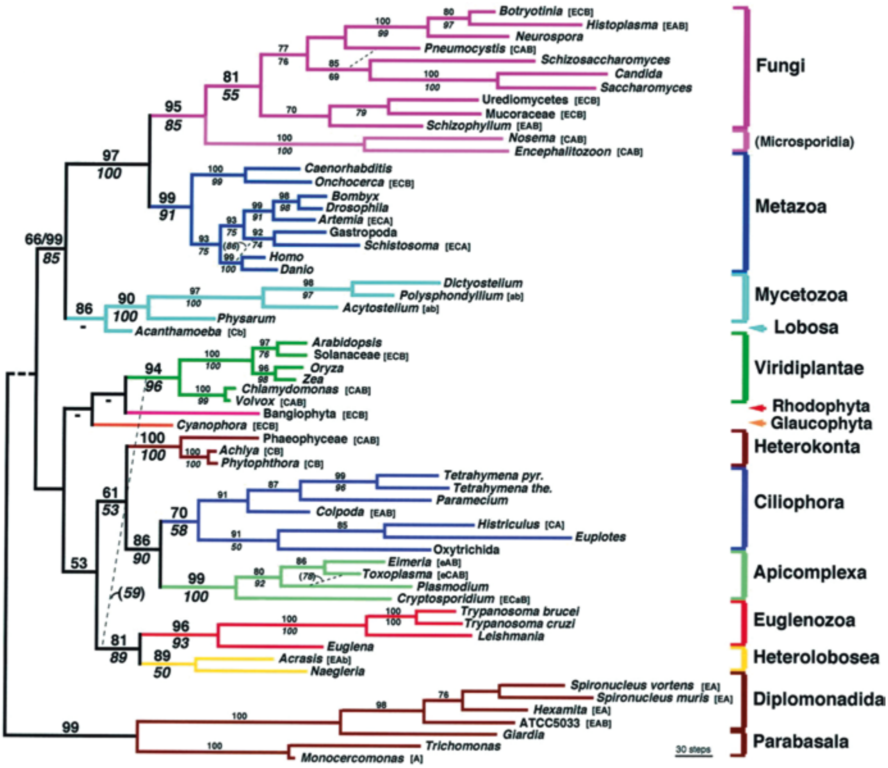


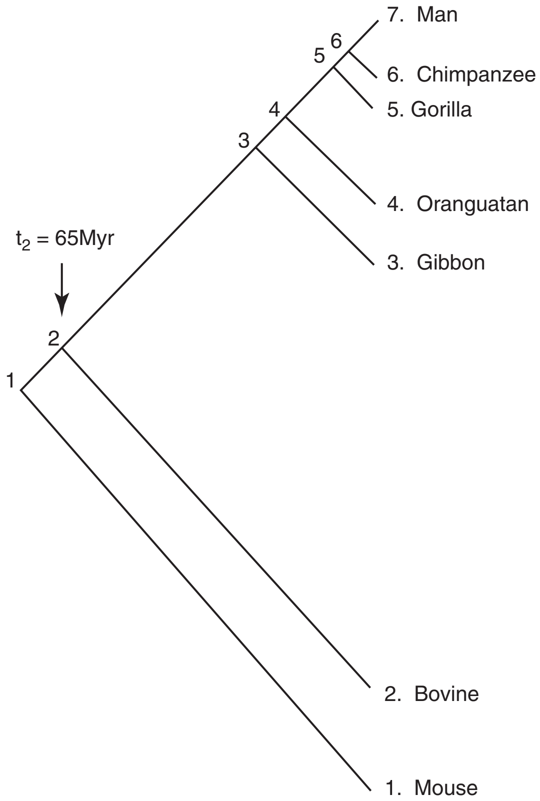
Fig. 10.10 The phylogenetic tree shows the 14 higher-order eukaryote taxa. (Source: Baldauf et al. 2000)

10.1.8 Evolution of Pathogen

Viruses are with high mutation rate and adapt quickly to environmental changes leading to the high genetic diversity. On the other hand, this fast evolution leaves behind significant marks in the genome of virus that can be connected with transmission dynamics and epidemiology. Evolutionary theory and sequence analysis played a role in understanding epidemiology of virus by figuring out the origin of time and geographical site of a virus. Analysis was able to provide information on transmission linkages or chains for a population.

Huet et al. (1990) inferred the origin and classified HIV into types, groups, and subtypes (Fig. 10.13). Epidemiological, physiological, and clinical evidences favored cross-species transmission of HIV from chimpanzee to humans (Castro-Nallar et al. 2012). Further, phylogenetic evidence corroborates this fact that HIV-1 and HIV-2 are due to several cross-species transmission events (Huet et al. 1990; Gao et al. 1992, 1999; Hahn et al. 2000; Plantier et al. 2009; Van Heuverswyn and Peeters 2007) (Fig. 10.14).

Fig. 10.11 The phylogenetic tree shows the divergence of apes and humans in vertebrates. (Source: Hasegawa et al. 1985)



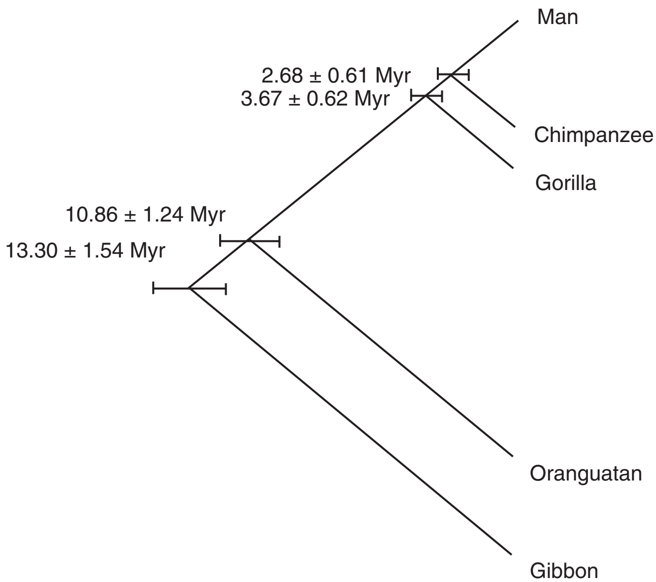


Fig. 10.12 The phylogenetic tree shows the divergence of humans from apes. (Source: Hasegawa et al. 1985)

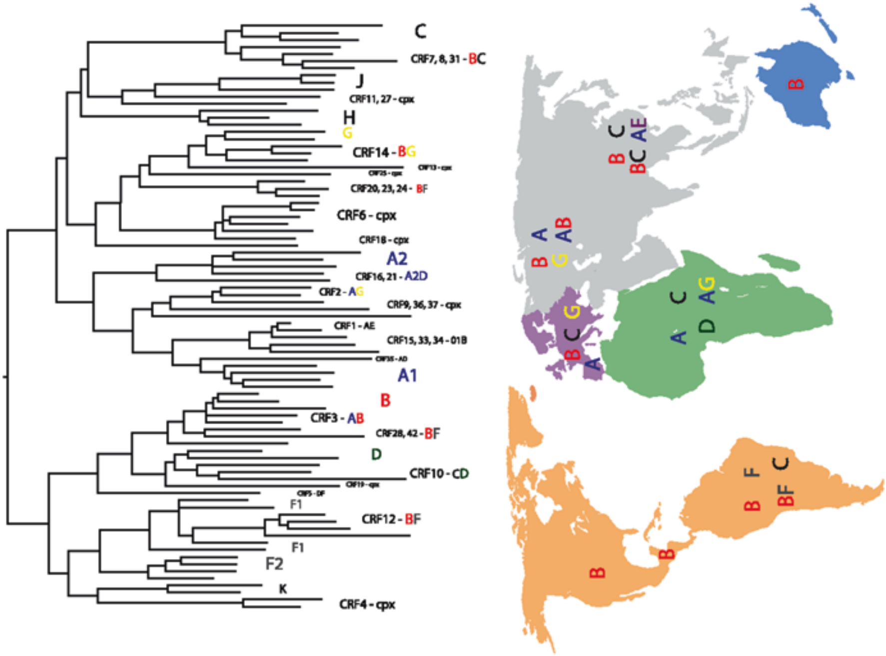


Fig. 10.13 Phylogenetic tree representation of HIV-1 and its subtypes. (Source: Castro-Nallar et al. 2012)

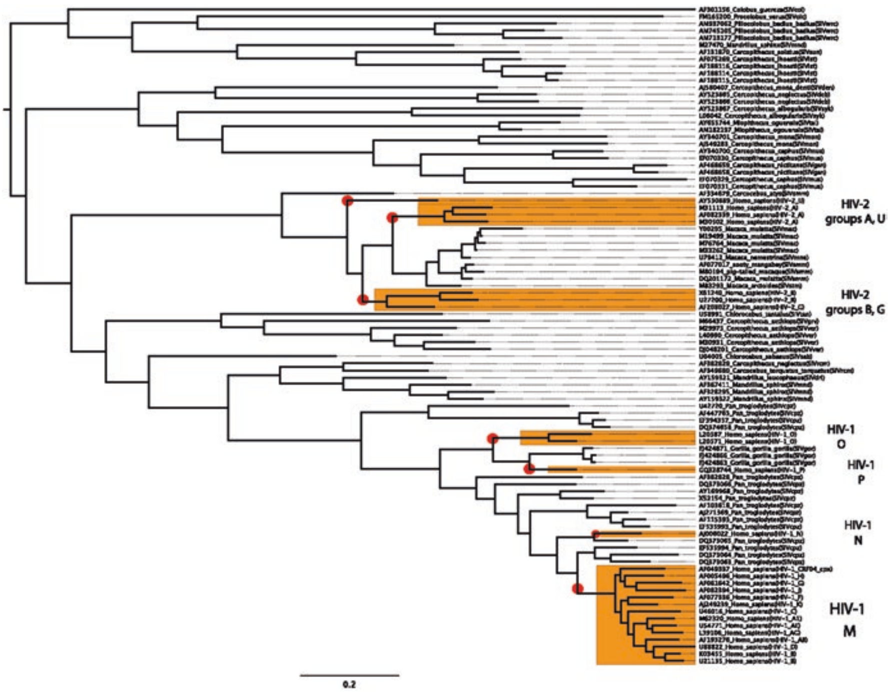


Fig. 10.14 Phylogenetic tree showing HIV cross-species transmission. (Source: Castro-Nallar et al. 2012)

Intensive studies were carried out on the evolution and divergence of HIV-1 and HIV-2 using phylogeny. The divergence time of HIV-1, HIV-2 (subtype A), and HIV-2 (subtype B) dated to the 1920s (Worobey et al. 2008), 1940 ± 16 (Lemey et al. 2003), and 1945 ± 14 (Lemey et al. 2003), respectively. Introduction of clade B of HIV-1 into North America dated to 1968 (1966–1970) (Gilbert et al. 2007; Pérez-Losada et al. 2010).

The emerging field of phylodynamics—"the melding of immunodynamics, epidemiology, and evolutionary biology …"—was used to understand the transmission dynamics, population dynamics, and within-host dynamics of virus or bacteria (Grenfell et al. 2004). Transmission dynamics helps in understanding diversity of an organism in transmission network constructed during a transmission event for potential therapy development. Population dynamics increases our understanding on patterns of diversity among populations throughout the length and breadth of infection, within host and transmission events. Within-host dynamics provide information on evolution of virus in the host which is associated with disease progression. There are two aspects within host dynamics which are observed in case of HIV. The first one is that evolution of HIV is different in specific tissues. It was revealed that HIV evolves at different rates in different compartments of the brain, which cannot be attributed to selective pressure, but can be related to viral expansion due to immune failure (Salemi et al. 2005). The second aspect is that HIV genetic diversity (variation) in the host leads to evolution of quasispecies (Holmes 2009). So, phylodynamics can be useful in relating epidemiological and evolutionary information which can be used for monitoring surveillance programs of a virus especially in case of HIV. Thus, phylogenetics can be used to identify evolution of virus in terms of origin, time of divergence, pathogen evolution, and understand phylodynamics.

10.2 Construction of Phylogenetic Trees

Data and tree construction methods used for construction of phylogenetic tree effect topology of the tree; therefore, it is worth to discuss on data and tree construction methods.

10.2.1 Data

Data generated via fingerprinting techniques such as rapid amplification polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), AFLP, SSCP, and sequence data (nucleotide and protein sequence data) are used for phylogeny. Data from fingerprinting techniques such RAPD, AFLP, and SSCP is converted to binary data (0/1). The "0s" represent the absence of band in the DNA fingerprinting techniques, whereas "1s" represent the presence of band in the DNA fingerprinting techniques. DNA or protein sequence data is generated by Sanger's method.

This binary or sequence data is either converted to distance or used directly in the form of character used to construct a phylogenetic tree. The fingerprinting data or the sequence data (DNA or protein) was known to influence the tree topology of the phylogenetic tree (Neelapu 2007; Devi et al. 2006, 2007; Padmavathi et al. 2003).

10.2.2 Tree-Constructing Methods

Broadly, there are two fundamental methods for constructing phylogenetic trees: distance or discrete character methods. Distance methods first convert data or aligned sequences into pairwise distance matrix. A correction is needed for these raw distances. These corrections are based on the assumptions of various substitution models proposed for both nucleic acid and protein sequence methods. A phylogenetic tree building method is then used to construct an evolutionary tree. Some of the tree-building methods are unweighted pair group method with arithmetic means (UPGMA), minimum evolution, neighbor joining, and Fitch-Margoliash.

UPGMA (Sokal and Michener 1958; Nei 1975) clusters data based on similarity and assumes that changes are accumulated at a constant rate among the lineages. In neighbor-joining method (Saitou and Nei 1987), a star tree in which terminal taxa are equidistant, is first established; then, two taxa are temporarily taken from the star to a new node, and the total distance in the new tree is recalculated; and the taxa are returned to the star and another pair of taxa is taken to repeat the operation. This process is continued until all the taxa are jointed in a completely resolved tree with the lowest total distance. In minimum evolution method (Takahashi and Nei 2000), the initial tree is created by clustering taxa using neighbor-joining method. Then, every possible tree is examined and one tree with minimum branch length is selected, thereby minimizing the total distance in a tree.

Discrete methods directly consider the state of each nucleotide or amino acid site in each sequence under comparison. The two discrete character methods are maximum likelihood and maximum parsimony. Maximum likelihood method (Cavalli-Sforza and Edwards 1967; Felsenstein 1973; Felsenstein 1981; Swofford et al. 1996) uses data to determine the probability of substitution, relative frequencies, and the different probabilities of transitions and transversions. It then selects the tree that maximizes the probability of good fit of the data. Maximum likelihood method presents an additional opportunity to evaluate trees with variations in mutation rates in different lineages; and also to use explicit evolutionary models such as the jukes-cantor and Kimura models.

Parsimony is another discrete character method that creates evolutionary trees based on a systematic search among possible trees for the fewest plausible mutational steps from a common ancestor necessary to account for two diverged lineages, and those trees that require the fewest changes are said to be most parsimonious (i.e., optimal) trees. The sum of the minimum possible substitutions over all sites is known as the tree length for that topology. The topology with the minimum tree length is known as the maximum parsimony tree. Three different types of searches the max-mini branch-and-bound search, min-mini heuristic search, and close-neighbor-interchange heuristic search are performed to generate maximum parsimony tree. The maximum parsimony method (Fitch 1971) produces many equally parsimonious trees. A majority-rule consensus method is used to produce a composite tree that is a consensus among all such trees.

10.2.3 Phylogeny Program Packages

All these clustering methods are available in various phylogenetic packages such as PHYLIP (Felsenstein 1989), PAUP (Swofford 1991), MEGA (Kumar et al. 2004), TreePuzzle (Schmidt et al. 2002), etc. (Table 10.1). The computational limits that were faced in running maximum parsimony and maximum likelihood method with increase in number of species and increase in length of the sequence in most packages are overcome in MEGA. Moreover, best tree editing options such as Tree Explorer program are available in MEGA, which makes phylogenetic inference from sequence data much easier.

10.3 Methods to Assess the Confidence of Phylogenetic Tree

The tree generated based on the input data and tree construction method is known as inferred tree. This inferred tree need not be the true tree for the given phylogenetic data. So, there is a requirement to test the reliability of the phylogenetic tree or portion of the tree. In methods like minimum evolution, maximum parsimony, and maximum likelihood, increase in tree number is observed as the sample size increases (Table 10.2). In these conditions, whether the tree is significant/better than another tree is to be confirmed. The reliability of the phylogenetic tree or portion of the tree is tested by sampling methods, whereas the significant difference of a tree over the other is confirmed by statistical tests.

10.3.1 Sampling Methods

The reliability of the phylogenetic tree or portion of the tree is tested by sampling methods such as bootstrapping, jackknifing, and Bayesian simulation.

10.3.1.1 Bootstrapping

Bootstrapping is random sampling with replacement of data (distances or sequence: nucleotide or protein) which addresses if any sampling errors occurred for the required analysis. In molecular phylogeny, bootstrapping repeatedly samples the

Table 10.1 Phylogenetic packages and resources available for phylogenetic analysis

S. No	Phylogeny packages	Description	Resource available at	References
1	**PHYL**ogeny **I**nference **P**ackage (PHYLIP)	PHYLIP is a free package available for inferring phylogenies. Molecular sequences; gene frequencies; restriction sites and fragments; distance matrices; and discrete characters are used as input. Parsimony, distance, and likelihood methods are used to construct phylogenetic tree	http://evolution.genetics.washington.edu/phylip/general.html	Felsenstein (1989)
2	**P**hylogenetic **A**nalysis **U**sing **P**arsimony (PAUP)	Phylogenetic relationships are performed for molecular, morphological, and behavioral data. The data for PAUP is NEXUS format. PAUP is used in the field of evolutionary biology, conservation biology, ecology, and forensic studies	http://paup.sc.fsu.edu/about.html	Swofford (1991)
3	MacClade's	MacClade is a phylogenetic analysis suite which is used for studying evolution of characters-based methods. The features include entering and editing data and phylogenies	http://macclade.org/index.html	Maddison and Maddison (1992)
4	FastDNAml	FastDNAml constructs phylogenetic tree for DNA sequences based on maximum likelihood method with improved performance	http://iubio.bio.indiana.edu/soft/molbio/evolve/fastdnaml/fastDNAml.html	Olsen et al. (1994)
5	**MOL**ecular **PHY**logenetics (Molphy)	MolPhy infers phylogeny of nucleic acid and protein using maximum likelihood	http://www.softpedia.com/get/Science-CAD/Molphy.shtml	Adachi and Hasegawa (1996)
6	GeneTree	GeneTree uses parsimony method to produce gene trees and species phylogenies	http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html	Page (1998)
7	Tree analysis using New Technology (TNT)	TNT is a parsimonious method used for searching parsimony tree	http://www.zmuc.dk/public/phylogeny/TNT	Goloboff (1999)
	TCS	TCS estimates gene genealogies within a population in a minimum spanning tree (rooted network tree) based on parsimony method. The accepted input data is sequence or absolute distances in NEXUS or PHYLIP format	http://darwin.uvigo.es/software/tcs.html	Clement et al. (2000)
8	GEODIS	GEODIS analyzes haplotypes in a genealogy to distinguish between historical divergence of populations and geographical separation	http://darwin.uvigo.es/software/geodis.html	Posada et al. (2000)

9	**P**hylogenetic **A**nalysis by **M**aximum **L**ikelihood (PAML)	PAML analyzes DNA or protein sequences using maximum likelihood	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang (2000)
10	PAL (Phylogenetic Analysis Library)	PAL facilitates construction of phylogenetic trees based on maximum likelihood, neighbor-joining, and least squares methods. PAL assesses the confidence of tree by bootstrapping, Kishino-Hasegawa-Templeton, and Shimodaira-Hasegawa tests. The divergence time in a tree can also be estimated	http://www.cebl.auckland.ac.nz/pal-project/	Drummond and Strimmer (2001)
11	Mr Bayes	Bayesian phylogenetic inference is based on Markov chain Monte Carlo (MCMC) methods	http://mrbayes.sourceforge.net/	Huelsenbeck and Ronquist (2001)
12	Data Analysis in Molecular Biology and Evolution (**DAMBE**)	DAMBE infers phylogeny from DNA and protein sequence by parsimony, distance, and likelihood methods. It can assess the confidence of tree either by bootstrapping or jackknifing. It can read and convert a number of file formats	http://dambe.bio.uottawa.ca/dambe.asp	Xia and Xie (2001)
13	TreePuzzle	TreePuzzle constructs phylogenetic tree from sequence data by maximum likelihood method. The confidence of the tree can be assessed by Kishino-Hasegawa test and Shimodaira-Hasegawa test	http://www.tree-puzzle.de/	Schmidt et al. (2002)
14	**M**olecular **E**volutionary **G**enetic **A**nalysis (MEGA)	Sophisticated and user-friendly software suite for sequence analysis, tree construction, and data visualization. Sequence Analyses suite includes phylogeny inference, dating and clocks, recognizing ancestral states, sequence alignment for DNA, and protein sequence data. The statistical methods maximum likelihood, distance methods, ordinary least squares, maximum parsimony, and composite likelihood are used for construction of tree. The powerful data visual tools are Alignment/Trace Editor, Tree Explorer, Data Explorers, Legend Generator, Gene Duplication Wizard, and Timetree Wizard	https://www.megasoftware.net/	Kumar et al. (2004)
15	IQPNNI	An efficient tree reconstruction method which is based on important quartet puzzling (IQP) and nearest neighbor interchange (NNI) operation	http://www.cibiv.at/software/iqpnni/	Vinh and von Haeseler (2004)
16	BayesPhylogenies	BayesPhylogenies is an application based on Bayesian Markov Chain Monte Carlo (MCMC) or Metropolis-coupled Markov chain Monte Carlo (MCMCMC) method which models the evolution	http://www.evolution.reading.ac.uk/BayesPhy.html	Pagel and Meade (2004)

(continued)

Table 10.1 (continued)

S. No	Phylogeny packages	Description	Resource available at	References
17	BAli-Phy	BAli-Phy is a phylogenetic application that generates multiple sequence alignments and evolutionary trees based on likelihood method from DNA and amino acid sequences	http://www.bali-phy.org/	Suchard and Redelings (2006)
18	BUCKy	BUCKy combines molecular data from multiple loci of sampled individuals to analyze the assumption of Bayesian concordance or discordance among gene trees	https://www.stat.wisc.edu/~ane/bucky/	Ané et al. (2007)
19	Bosque	Phylogenetics application for sequence management, alignment, and construction of phylogenetic tree	http://bosque.udec.cl/	Ramírez and Ulloa (2008)
20	FastTree	FastTree infers phylogenetic trees from nucleotide or protein sequences using maximum-likelihood method 1000 times faster than regular maximum-likelihood applications with reasonable time and memory	http://www.microbesonline.org/fasttree/	Price et al. (2009)
21	MetaPIGA	MetaPIGA infers phylogeny under maximum likelihood for binary and character data (nucleotide and protein)	http://www.metapiga.org/	Raphaël and Milinkovitch (2010)
22	Mesquite	Mesquite is modular software for evolutionary biology to help biologists in population genetics, phylogenetic analysis, and geological timescale analysis	http://mesquiteproject.org/	Maddison and Maddison (2011)
23	**B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees (BEAST)	Bayesian analysis based on MCMC method infers phylogenetic trees	http://beast.community/	Drummond et al. (2012)
24	Armadillo	Armadillo is an open-source computer applications suite for bioinformatics analysis. The following bioinformatics tasks are performed using Armadillo—"automatic BLAST queries, multiple sequence alignment, evolutionary model inference, construction of phylogenetic trees (distance, maximum likelihood, maximum parsimony, and Bayesian methods), visualization of phylogenetic trees and networks, phylogenetic tree manipulation, and horizontal gene transfer detection"	http://www.bioinfo.uqam.ca/armadillo/	Lord et al. (2012)
25	IQ-TREE	A fast and efficient maximum likelihood method to infer phylogenetic tree from the data	http://www.iqtree.org/	Nguyen et al. (2015)
26	TOPALi	TOPALi estimates phylogenetic tree based on Bayesian analysis (BA) and maximum likelihood (ML)	http://www.topali.org/	McGuire and Wright (2000)

Table 10.2 The number of rooted trees and unrooted trees for *n* sequences

Number of taxa	Number of unrooted trees	Number of rooted trees
	Formula	
	$\text{Nu} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$	$\text{Nr} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

data to construct the phylogenetic tree and gives us the chance to assess the strength of the original tree. If the data resampling generates different trees when compared with the original tree, then the tree topology is based on the data with weak phylogenetic signals. If the data resampling generates tree similar to the original tree, then the tree topology is based on the data with enough phylogenetic signals. Thus, bootstrapping (resampling data) provides insights on the confidence of the tree topology.

Two types of bootstrapping are used in phylogenetic analysis: parametric or nonparametric bootstrapping. If the data is disturbed by random sampling generating new dataset, then it is nonparametric bootstrapping. If the data is disturbed by particular order to generate new dataset, then it is parametric bootstrapping. The other types of bootstrapping are case resampling, Bayesian bootstrap, smooth bootstrap, resampling residuals, Gaussian process regression bootstrap, wild bootstrap, and block bootstrap (time series: simple block bootstrap, time series: moving block bootstrap, cluster data: block bootstrap).

If bootstrapping is repeated 100–1000 times or even more to reconstruct phylogenetic trees, then certain parts of the tree have different topology when compared with the original inferred tree. All these bootstrapped trees are summed up into a consensus tree based on a majority rule. The most supported branching patterns shown at each node are labeled with bootstrap values. Thus, bootstrap offers a measure for estimating the confidence levels of the tree topology.

10.3.1.2 Jackknifing

Jackknifing is another resampling technique where half of the dataset is randomly deleted, generating datasets half-original. Initially, a phylogenetic tree is constructed with the original dataset, then with each new dataset generated by jackknifing, a phylogenetic tree is constructed using the same method as the original. Sampling generates different trees when phylogenetic signals are weak, whereas sampling generates similar tree when phylogenetic signals are strong. Thus, jackknifing (resampling data) can also be used to assess the confidence of the tree topology.

10.3.1.3 Bayesian Method

Bayesian method based on MCMC approach resamples data thousands or millions of steps or iterations. The sample datasets are used to reconstruct phylogenetic trees similar to original inferred tree. The posterior probabilities designated at each intersection of a best Bayesian tree measure the confidence levels of the tree topology.

10.3.2 Statistical Methods

The significant difference of a tree over the other is confirmed by statistical tests such as Kishino-Hasegawa Test and Shimodaira-Hasegawa Test.

10.3.2.1 Kishino-Hasegawa Test

Kishino-Hasegawa (KH) test compares two tree topologies to differentiate one tree over the other (Kishino and Hasegawa 1989). Though KH test can be used for differentiating trees generated through methods such as distance, parsimony, and likelihood, Kishino-Hasegawa developed this test specifically for parsimonious trees. The KH test (statistical method) is paired Student *t*-test based on null hypothesis that the "two competing tree topologies are not significantly different…." The standard deviation of the difference between branch lengths at each informative site between two trees is estimated. Then the derived *t*-value is compared with the *t*-distribution values either to accept or reject the null hypothesis at certain significant levels (with probability e.g., $P < 0.05$).

$$t = \frac{\mathrm{Da} - \mathrm{Dt}}{\mathrm{SD} / \sqrt{n}}$$

$\mathrm{df} = (n - 1)$ where t is the test statistical value, Da is the average site-to-site difference between the two trees, Dt is the total difference of branch lengths of the two trees, SD is the standard deviation, n is the number of informative sites, and df is the degree of freedom.

10.3.2.2 Shimodaira-Hasegawa Test

Shimodaira-Hasegawa (SH) developed a statistical test for ML trees based on likelihood ratio using the $\chi 2$ test to estimate the goodness of fit of two competing trees (Shimodaira and Hasegawa 1999). The log likelihood scores lnLA and lnLB for tree

A and tree B are obtained first, for the two competing trees. Then the log ratio of the two scores is obtained by $d = 2(\ln LA - \ln LB) = 2 \ln (LA/LB)$ and used to test against the $\chi 2$ distribution from a table. The resulting probability value (*P*-value) determines whether the difference between the two trees is significant or nonsignificant.

10.4 Conclusion

Molecular phylogeny establishes the relationships among the set of objects in the study. Binary data ("0"/"1") from RAPD, RFLP, AFLP, SSCP, and sequence data (DNA or protein) from the set of objects are used to construct phylogenetic tree. The different tree construction methods are UPGMA, NJ, ME, FM, MP, and ML. Molecular phylogeny has a wide range of applications and if the interpretation of the evolutionary patterns is not appropriate, then the inference of the study may be misleading. The interpretation of the tree is always dependent on assessing the confidence of the phylogenetic tree. Sampling methods (bootstrapping, jackknifing, and Bayesian simulation) and statistical methods (KH test and SH test) can be used to assess the confidence of the phylogenetic tree. Thus, if the confidence of the phylogenetic tree generated is good, then the interpretation or inference of the study will not be misleading.

Acknowledgment The authors are grateful to Gandhi Institute of Technology and Management (GITAM) Deemed-to-be-University, for providing necessary facilities to carry out the research work and for extending constant support in writing this review.

References

Adachi J, Hasegawa M (1996) Molphy, version 2.3. Programs for molecular phylogenetics based on maximum likelihood. In: Ishiguro M, Kitagawa G, Ogata Y, Takagi H, Tamura Y, Tsuchiya T (eds) Computer science monographs. Institute of Statistical Mathematics, Tokyo

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. Mol Biol Evol 24(2):412–426

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290(5493):972–977

Castro-Nallar E, Perez-Losada M, Burton GF, Crandall KA (2012) The evolution of HIV: inferences using phylogenetics. Mol Phylogenet Evol 62:777–792

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. Mol Ecol 9:1657–1660

Devi KU, Reineke A, Reddy NNR, Rao CUM, Padmavathi J (2006) Genetic diversity, reproductive biology, and speciation in the entomopathogenic fungus *Beauveria bassiana* (Balsamo) Vuillemin. Genome 49(5):495–504

Devi UK, Reineke A, Rao UCM, Reddy NRN, Khan APA (2007) AFLP and single-strand conformation polymorphism studies of recombination in the entomopathogenic fungus *Nomuraea rileyi*. Mycol Res 111(6):716–725

Drummond A, Strimmer K (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17:662–663
Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973
Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249
Felsenstein J (1981) Evolutionary trees from gene-frequencies and quantitative characters – finding maximum-likelihood estimates. Evolution 35:1229–1242
Felsenstein J (1989) PHYLIP – phylogeny inference package (version 3.2). Cladistics 5:164–166
Fitch WM (1971) Towards defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416
Frech C, Chen N (2010) Genome-wide comparative gene family classification. PLoS One 5(10):e13409. https://doi.org/10.1371/journal.pone.0013409
Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, Greene BM, Sharp PM, Shaw GM, Hahn BH (1992) Human infection by genetically diverse SIVSM-related HIV-2 in West Africa. Nature 358:495–499
Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH (1999) Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature 397:436–441
Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M (2007) The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566–18570
Goloboff PA (1999) Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15:415–428
Grenfell B, Pybus O, Gog J, Wood J, Daly J (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332
Hahn BH, Shaw GM, De Cock KM, Sharp PM (2000) AIDS as a zoonosis: scientific and public health implications. Science 287:607–614
Hardison RC (2012) Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med 2(12):a011627. https://doi.org/10.1101/cshperspect.a011627
Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2):160–174
Holmes EC (2009) The evolution and emergence of RNA viruses. Oxford University Press, New York
Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755
Huet T, Cheynier R, Meyerhans A, Roelants G, Wain-Hobson S (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature 345:356–359
Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179
Kumar S, Tamura K, Nei M (2004) MEGA3: an integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163
Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme A-M (2003) Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci U S A 100:6588–6592
Lord E, Leclercq M, Boc A, Diallo AB, Makarenkov V (2012) Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. PLoS One 7(1):e29903. https://doi.org/10.1371/journal.pone.002990
Maddison WP, Maddison DR (1992) MacClade. Sinauer Associates, Sunderland
Maddison WP, Maddison DR (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org
Maeshima M (2000) Vacuolar H^+-pyrophosphatase. Biochim Biophys Acta 1465:37–51
Margos G, Vollmer SA, Ogden NH, Fish D (2011) Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. Infect Genet Evol 11(7):1545–1563

McGuire G, Wright F (2000) TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics 16(2):130–134

Neelapu NRR (2007) Investigation on existence and mechanism of recombination and molecular phylogeny of mitosporic entomopathogenic fungi *Beauveria bassiana* (Balsamo) Vuillemin and *Nomuraea rileyi* (Farlow) Samson. Doctoral dissertation, Andhra University, Visakhapatnam, India

Neelapu NRR, Reineke A, Chanchala UMR, Koduru UD (2009) Molecular phylogeny of asexual entomopathogenic fungi with special reference to *Beauveria bassiana* and *Nomuraea rileyi*. Rev Iberoam Micol 26(2):129–145

Nei M (1975) Molecular population genetics and evolution. North-Holland, Amsterdam

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274

Olsen GJ, Matsuda H, Hagstrom R, Overbeek R (1994) FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Bioinformatics 10(1):41–48

Opazo JC, Homan FG, Storz JF (2008) Genomic evidence for independent origins of like globin genes in monotremes and therian mammals. Proc Natl Acad Sci U S A 105:1590–1595

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740

Padmavathi J, Uma Devi K, Rao CUM, Reddy NNR (2003) Telomere fingerprinting for assessing chromosome number, isolating typing and recombination in the entomopathogen *Beauveria bassiana*. Mycol Res 107(5):572–580

Page RDM (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819–820

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581

Pérez-Losada M, Jobes DV, Sinangil F, Crandall KA, Posada D, Berman PW (2010) Phylodynamics of HIV-1 from a phase-III AIDS vaccine trial in North America. Mol Biol Evol 27:417–425

Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, Damond F, Robertson DL, Simon F (2009) A new human immunodeficiency virus derived from gorillas. Nat Med 15:871–872

Posada D, Crandall KA, Templeton AR (2000) GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Mol Ecol 9:487–488

Pozio E, Hoberg E, La Rosa G, Zarlenga DS (2009) Molecular taxonomy, phylogeny and biogeography of nematodes belonging to the *Trichinella genus*. Infect Genet Evol 9(4):606–616

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650

Ramírez-Flandes S, Ulloa O (2008) Bosque: integrated phylogenetic analysis software. Bioinformatics 24(21):2539–2541

Raphaël H, Milinkovitch MC (2010) MetaPIGA v2.0: maximum likelihood large phylogeny estimation using the metapopulation genetic algorithm and other stochastic heuristics. BMC Bioinforma 11:379

Rea PA, Kim Y, Sarafian V, Poole RJ, Davies JM, Sanders D (1992) Vacuolar H^+-translocating pyrophosphatase: a new category of ion translocase. Trends Biochem Sci 17(9):348–352

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM, McGrath MS (2005) Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. J Virol 79:11343–11352

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. J Univ Kans Sci Bull 28:1409–1438
Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048
Suneetha G, Neelapu NRR, Surekha C (2016) Plant vacuolar proton pyrophosphatases (VPPases): structure, function and mode of action. Int J Recent Sci Res 7(6):12148–12152
Swofford DL (1991) PAUP: Phylogenetic Analysis Using Parsimony, version 3.1 Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer, Sunderland
Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17:1251–1258
Teugels G (1996) Taxonomy, phylogeny and biogeography of catfishes (*Ostariophysi*, *Siluroidei*): an overview. Aquat Living Resour 9(S1):9–34. https://doi.org/10.1051/alr:1996039
Thompson RCA (2008) The taxonomy, phylogeny and transmission of *Echinococcus*. Exp Parasitol 119(4):439–446
Van Heuverswyn F, Peeters M (2007) The origins of HIV and implications for the global epidemic. Curr Infect Dis Rep 9:338–346
Vinh LS, von Haeseler A (2004) IQPNNI: moving fast through tree space and stopping in time. Mol Biol Evol 21(8):1565–1571
Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
Woese CR, Fox GE (1997) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090
Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576–4579
Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe J-J, Kabongo J-MM, Kalengayi RM, Van Marck E, Gilbert MTP, Wolinsky SM (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661–664
Xia X, Xie Z (2001) DAMBE: data analysis in molecular biology and evolution. J Hered 92:371–373
Yang Z (2000) Phylogenetic analysis by maximum likelihood (PAML). University College, London