# Phylogenetic Analyses

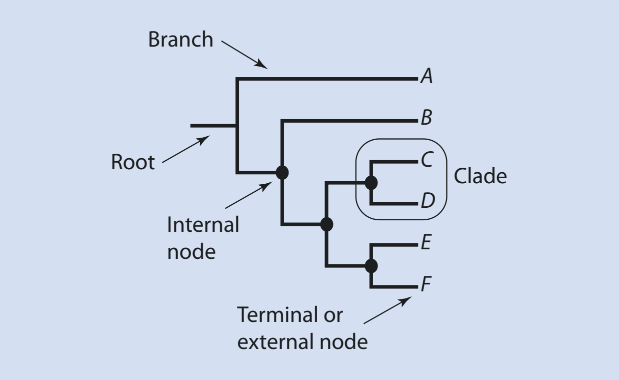
C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_8

- Phylogenetic trees represent the evolutionary relationships of sequences or species (or other taxonomic units) and can be shown as phylograms, cladograms; ultrametric trees, or unrooted.
- Networks or consensus methods can be used to summarize the information of multiple trees.
- Many tree building methods rely on explicit models of sequence evolution; models of nucleotide substitution are nested and can be derived from a general model, whereas amino acid models are broadly classified into empirical and mechanistic models.
- The most widely applied methods of phylogenetic reconstruction are neighbour joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI).
- Support of branches within a phylogenetic tree can be (among others) measured by bootstrapping or likelihood ratio test approaches.

## 8.1 Trees

The main aim of phylogenetic systematics is the reconstruction of evolutionary relationships which are represented by a tree. In a phylogenetic tree, the pattern of branches connected by internal nodes (topology) illustrates the relationships of the included terminals (◘ Fig. 8.1). In ◘ Fig. 8.1, the terminals represent the taxa A to F. When describing the topology of a tree, it has to be kept in mind that rotation of the axis of internal nodes does not change the topological information (◘ Fig. 8.2). A handy way to describe trees is to refer to clades (◘ Fig. 8.1), which are monophyletic units comprising an ancestor (internal node) and all of its descendants. For example, taxa C and D form a monophyletic group in ◘ Fig. 8.1. Moreover, referring to sister lineages (or groups) is a good way to describe trees. For example, taxon C is the sister lineage (or group) of taxon D in ◘ Fig. 8.1. Often misused when describing trees is the term «basal» (Krell and Cranston 2004), e.g. when referring in ◘ Fig. 8.1 to taxon A as «basal». This is wrong, as basal would imply that taxon A is the ancestral group – which obviously cannot be correct for an extant taxon. However,

◘ **Fig. 8.1** Terms describing the topology of a phylogenetic tree



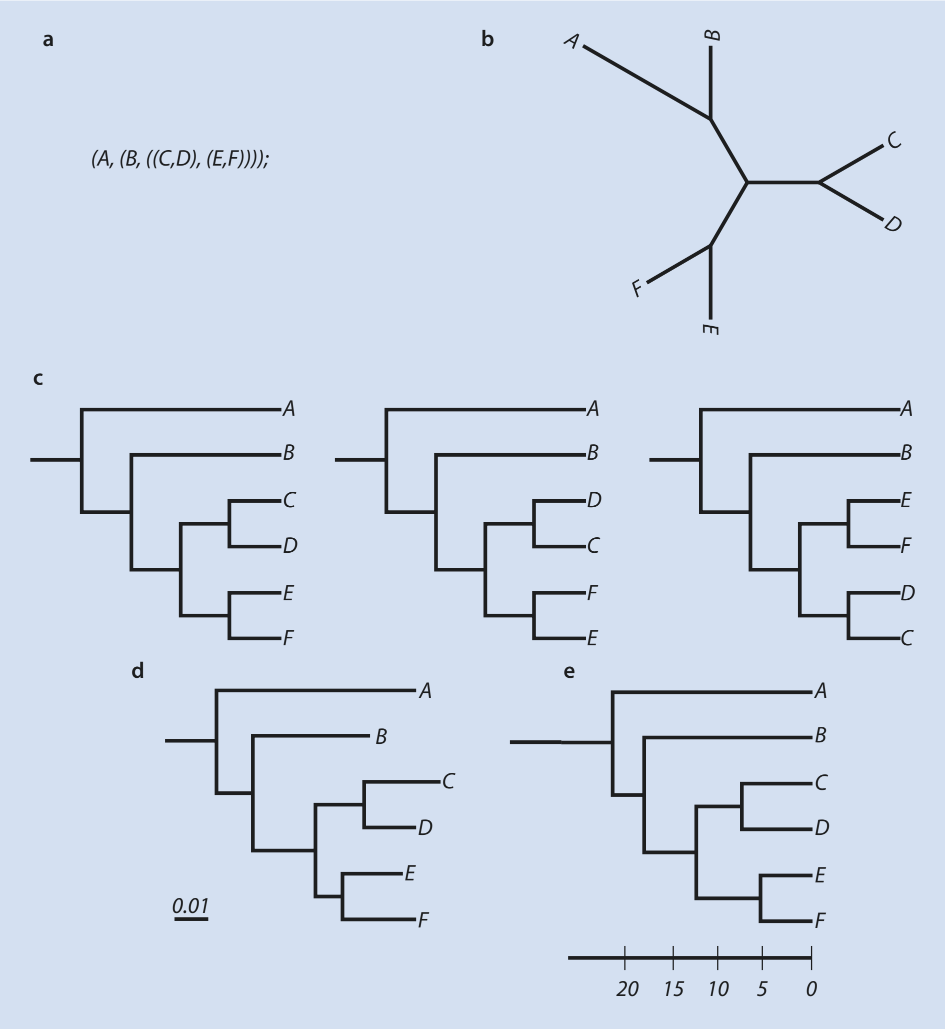


**Fig. 8.2** Different ways to represent the same tree topology. **a** Newick format. **b** Unrooted tree. **c** Three different representations of the same topology as a cladogram. **d** Phylogram. **e** Ultrametric tree with time axis

taxon A could be described as «basally branching», as the branch leading to this terminal is closest to the root.

One major problem for phylogenetic analyses is the giant number of possible tree topologies with increasing number of taxa. The number of possible strictly bifurcating unrooted trees ($N_u$) can be calculated as follows:

$$N_u = (2n-5)! \div 2^{n-3}(n-3)! \tag{8.1}$$

In this formula, $n$ denotes the number of terminals (e.g. taxa, OTUs, sequences) included in the tree. For example, for 4 taxa, the number of unrooted tree topologies is 3, whereas for 10 taxa, there are already 2,027,025 different topologies. The number of possibilities grows exponentially with the number of included terminals. When including 60 terminals, there are more possible tree topologies ($\sim 10^{94}$) than atoms in the universe ($\sim 10^{82}$). This sheer incomprehensible large number of possibilities is a major problem for all phylogenetic methods which include steps analysing all of them.

There are different ways to represent the topology of a tree (◘ Fig. 8.2). The standard output format of most phylogenetic software is the Newick format, where nested relationships are shown using brackets (◘ Fig. 8.2a). This format can easily be translated in an unrooted tree (◘ Fig. 8.2b). Unrooted trees can be polarized by choosing outgroups (Nixon and Carpenter 1993). This choice usually depends on prior knowledge, and correct outgroup choice is crucial for every analyses. If outgroups are unknown or not included in the analyses, midpoint rooting can be alternatively used for rooting. In this case, the root is placed at the midpoint of the longest distance between two terminals in a tree. However, midpoint rooting assumes that all included terminals evolve at the same rate and may fail to place the root correctly when this assumption is violated (Hess and De Moraes Russo 2007). Finally, the root of the tree could be inferred as part of the phylogenetic analysis when using nonreversible models of sequence evolution (see below), where different placements of the root affect the outcome of the analysis (Huelsenbeck et al. 2002a). If trees are represented as cladograms (◘ Fig. 8.2c), they only contain topological information. However, if they are represented as phylograms (◘ Fig. 8.2d) or ultrametric trees (◘ Fig. 8.1e), the length of branches carries additional information. In phylograms, branch length is proportional to evolutionary change. A typical measure of branch lengths for molecular data is the average number of substitutions per site in the alignment. The sum of the lengths of the branches linking two terminals (but also internal nodes) in a phylogram is called patristic distance (Fourment and Gibbs 2006). Ultrametric trees are reconstructed under the assumption that the change indicated by the length of branches is proportional to time («molecular clock», ► see also Sect. 8.7). Terminals in an ultrametric tree are equidistant from the root, which means that all paths of branches leading from the root to terminals have the same length. Phylogenetic divergence times can be estimated by calibrating ultrametric trees using palaeontological or biogeographic data (Donoghue and Benton 2007; Heads 2005). There are several applications available to visualize trees, and among the most widely used ones are DENDROSCOPE (Huson et al. 2007), ETE (Huerta-Cepas et al. 2010), FIGTREE (► http://tree.bio.ed.ac.uk/software/figtree/), ITOL (Letunic and Bork 2016) and TREEVIEW (Page 1996).

Sometimes, it is desirable to summarize the information of two or more topologies in a consensus tree. Several methods for building consensus trees are available (Wilkinson 1994), but only two of them are widely used in phylogenetic systematics: strict consensus and majority-rule consensus (◘ Fig. 8.3). In a strict consensus, only those internal nodes that can be found in all summarized topologies are displayed (◘ Fig. 8.3b); all other nodes are collapsed into multifurcations. Majority-rule consensus trees show those internal nodes which are found in more than half of all summarized topologies (◘ Fig. 8.3c); nodes that do not fulfil this criterion are collapsed. Usually, the frequency how often a node appears is indicated in the tree. Majority-rule consensus trees are widely used to summarize trees from bootstrap analyses and Bayesian inference (► see Sect. 8.6). Finally, there is a set of methods that derive a phylogenetic hypothesis (tree topology) from combining the topological information of different source trees. This so-called supertree approach differs from consensus methods, as it does not need an identical set of terminals to combine

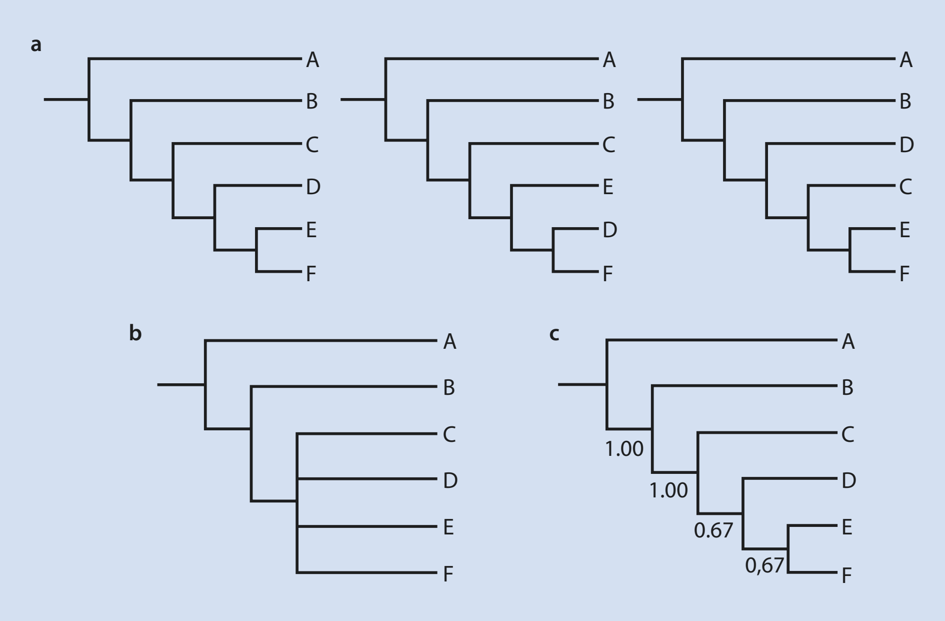


**Fig. 8.3** Consensus tree methods. **a** Cladograms of three different topologies. **b** Strict consensus, summarizing those nodes found in all trees. **c** Majority-rule consensus, summarizing those nodes which are found in more than 50% of the trees. Frequency of the occurrence of nodes is given at the nodes

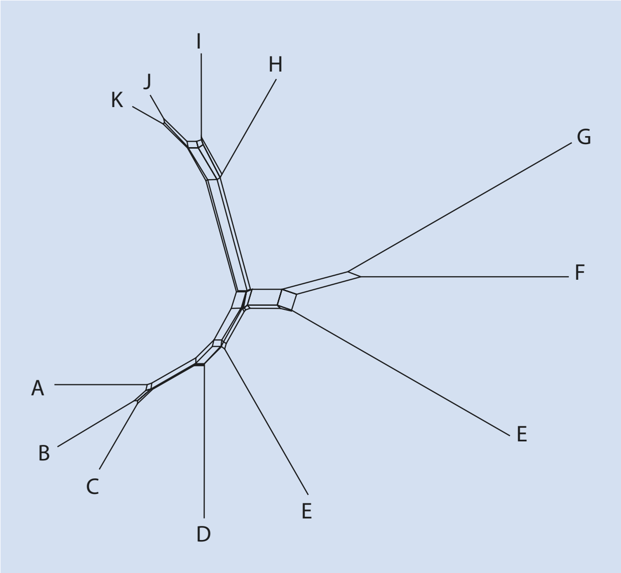
trees. Instead, supertrees can be built from topologies which overlap in its included terminals (Bininda-Emonds 2004). Even though supertrees can be build using a modified version of the strict consensus, by far the most widespread attempt is using a form of matrix representation (Baum 1992; Ragan 1992). Under this approach, all internal nodes of the input trees are coded as characters in a matrix. Each terminal which appears in at least one of the trees will be included in the matrix, and either are coded as present (1) or absent (0) for each character. Finally, distance or parsimony methods (► see Sect. 8.5) can be used to reconstruct the supertree.

Every tree is a special kind of a graph. A graph can be broadly defined as a representation of a finite number of nodes connected by branches (edges) to show their relationships (Huson et al. 2010). Trees are connected graphs without cycles, which means there are no reticulations. However, phylogenetic trees might not always be the best way to represent evolutionary relationships (Doolittle and Bapteste 2007). For example, under the presence of hybridization, horizontal gene transfer or recombination reticulate relationships between nodes should be assumed. In this case, networks (◘ Fig. 8.4), which are connected graphs with cycles, are a better way to illustrate evolutionary relationships (Huson and Bryant 2006; Posada and Crandall 2001). Such networks can also be used to visualize conflict within phylogenetic datasets.

## 8.2 Models of Nucleotide Substitution

Under the assumption of a constant evolutionary rate over time, a linear increase of the number of nucleotide substitutions should be expected after divergence of a pair of sequences. However, as there might be back substitutions, multiple substitutions or

**Fig. 8.4** Example of a phylogenetic network



convergent substitutions, comparison of observed distances (p-distances) between pairs of sequences will show a level of saturation after some time of divergence (Page and Holmes 1998). To correct for this saturation, probabilistic models of sequence evolution are used to calculate expected distances. Most methods for phylogenetic reconstruction rely on explicitly formulated models of sequence evolution. Such models are incorporated within distance methods, maximum likelihood and Bayesian inference (▶ see Sect. 8.5). Nucleotide substitution models in use for phylogenetic inference make several assumptions to model substitutions as a stochastic process:

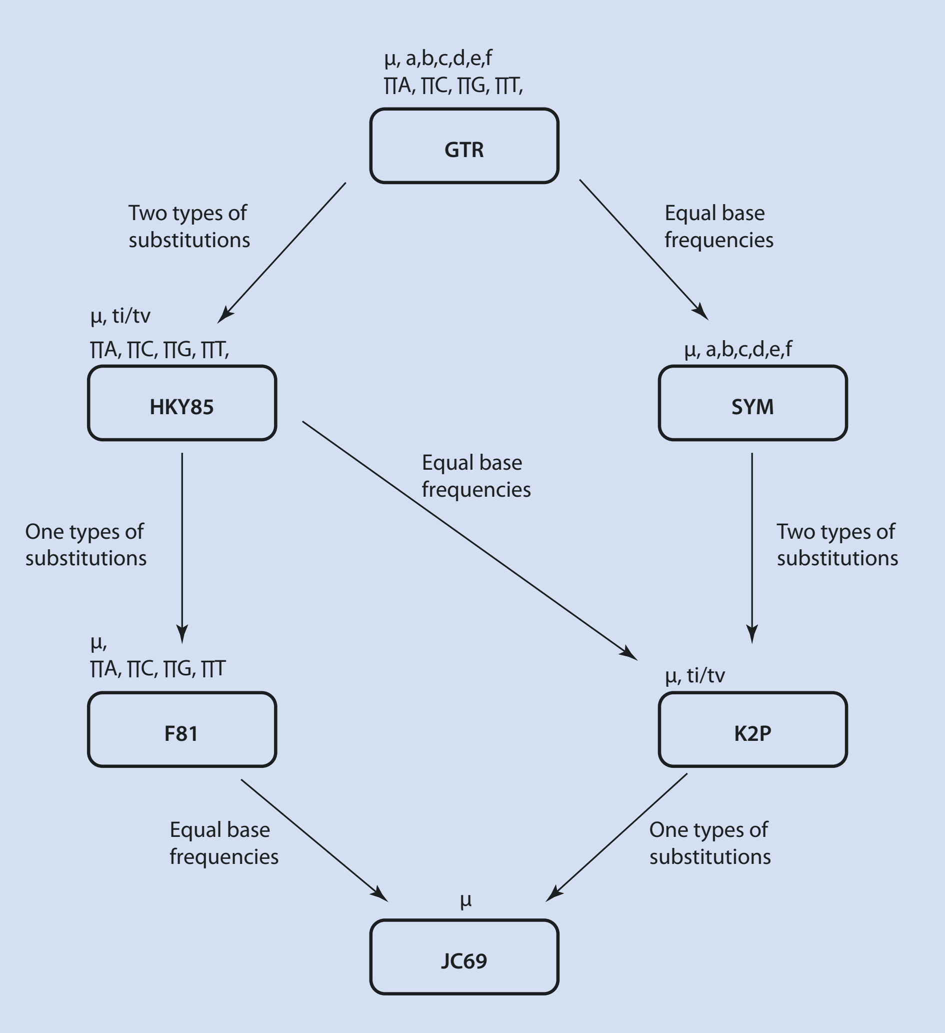
I. For every site of a sequence, it is assumed that the rate of change from one base to another is independent from the history of this site (Markov property).
II. It is assumed that substitution rates are not changing over time (homogeneity).
III. Equilibrium of base frequencies is assumed (stationarity).

Models that fit this description are called time-homogeneous time-continuous stationary Markov models. Substitution rates are summarized by such models in a rate matrix (or Q-matrix), where each entry specifies the probability for any possible nucleotide substitution. Usually, models used in molecular phylogenetics are time reversible, thereby additionally assuming that the rate of change from one base i to another base j is identical to the rate of a change from j to i (j and i can be all possible bases, but must be different bases). The most general model of nucleotide substitutions is the general time reversible (GTR) model (Rodríguez et al. 1990; Tavare 1986), which is summarized by the following Q-matrix:

$$Q = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix} \tag{8.2}$$

Formula 8.2: Q-matrix of the GTR model. Herein, $\mu$ is the overall substitution rate, and $\pi_X$ refers to the different base frequencies (with $X$ either A, C, G or T). The letters $a$ to $f$ represent the frequency of possible substitutions (e.g. $a$ is the frequency of substitutions from A to C or C to A).

The GTR model infers parameters for six different (reversible) substitution types (A–C, A–G, A–T, C–G, C–T, G–T), overall substitution rate and base frequency from the underlying data (a sequence alignment). For the rate of substitution types, one of these parameters is set to 1, whereas all other parameters are relative to the fixed parameter. Thereby in this case, there are five free parameters for the substitution types. There are several other models in use (▫ Fig. 8.5), which can be derived from the GTR model by



▫ **Fig. 8.5** Models of nucleotide substitutions and their interrelationships. Models are nested and can be derived from the general time reversible (GTR) model by restricting parameters. Open parameters are given above the boxes for each model. Herein $\mu$ is the overall substitution rate; $\pi_X$ refers to the different base frequencies (with X either A, C, G or T). The letters a to f represent the frequency of possible substitutions, whereas ti/tv represents the frequency of transitions to transversions. The restricted parameter to transform a more general model to a more restricted one is given at the *arrows* restricting some of the parameters. For example, in the HKY85 model (Hasegawa et al. 1985), only two types of substitutions (transitions vs. transversions) are distinguished. In the K2P model (Kimura 1980) also, only these two substitution types are distinguished, but additionally equal base frequencies (0.25 for each base) are assumed. The most restricted (and historically oldest) model is the JC69 model (Jukes and Cantor 1969), where all substitution types are assumed to be equally probable and base frequencies are fixed to be equal. A detailed description (including Q-matrices) of models for nucleotide substitutions can be found in Yang (2006) and Page and Holmes (1998).

All discussed models assume that the evolutionary rate is the same for every position of the sequence alignment. However, this assumption has been shown to be unrealistic when working with real data. As such the mutation rate can vary among bases. For example, G and C nucleotides are twice as mutable than A and T nucleotides in most species across the tree of life (Hodgkinson and Eyre-Walker 2011). This is due to the effect that in CpG dinucleotides (a C followed by a G) cytosines are often methylated and thereby prone to deamination, resulting in a T nucleotide (Fryxell and Moon 2005). For most datasets, the rate of fixation of mutations also varies among sites, due to the effect of different selective pressures. Obvious examples are protein-coding genes, where the codon positions evolve under different rates, with the third position usually accumulating substitutions much faster than the other two positions. Likewise, different selective pressures act on different regions of ribosomal RNA genes, and usually conserved and variable regions can be distinguished. By ignoring these variations across sites, the expected distance between a pair of sequence will be underestimated. To include rate heterogeneity across alignment sites, a statistical distribution is used to allow different sites to fall into categories of different substitution rates. Usually, a «gamma model» is used, with several categories of rates approximating a gamma distribution (Yang 1994). The shape of the gamma distribution is defined by the parameter $\alpha$ (◘ Fig. 8.6), which has to be determined to fit the gamma model for a given dataset. Typically, the shape parameter α is rather small (<1) (Yang 1996) resulting in a skewed L-distribution, reflecting that most of the sites show low substitution rates (or are invariable), whereas few sites range in a spectrum from low to high substation rates (Yang 2006). Large values of $\alpha$ would result in a rather bell-shaped distribution where most sites evolve under a similar rate. The continuous gamma distribution can be divided into categories of equal probability. Based on a comparison of several datasets, the use of six to ten rate categories has been suggested as a good approximation of rate heterogeneity (Jia et al. 2014). Models of sequence evolution which incorporate the gamma distribution are marked with a «+ Γ» or «+ G». Inclusion of a gamma distribution is computationally time and memory intensive, which can be a problem for large-scale phylogenomic analyses. Stamatakis (2006) developed a method to approximate rate heterogeneity called «CAT model» which is computationally much faster than inferring the gamma distribution. This approach is implemented in the popular software RAXML (Stamatakis 2014).

Another modification to account for rate heterogeneity is the incorporation of the proportion of invariant sites into models of sequence evolution (Fitch and Margoliash 1967b). Models using this modification are marked with a «+ I». If all alignment sites would change at the same rate, as assumed by all models discussed here, the number of substitutions should follow a Poisson distribution. Real datasets usually do not fit this distribution. However, the exclusion of invariant sites allows a better fit. Models including both modifications (+ I + Γ) assume that a proportion of sites are invariable, while the rates of the remaining sites are gamma distributed (Gu et al. 1995). It is discussed if such kind of models should be used at all, given that the amount of invariable

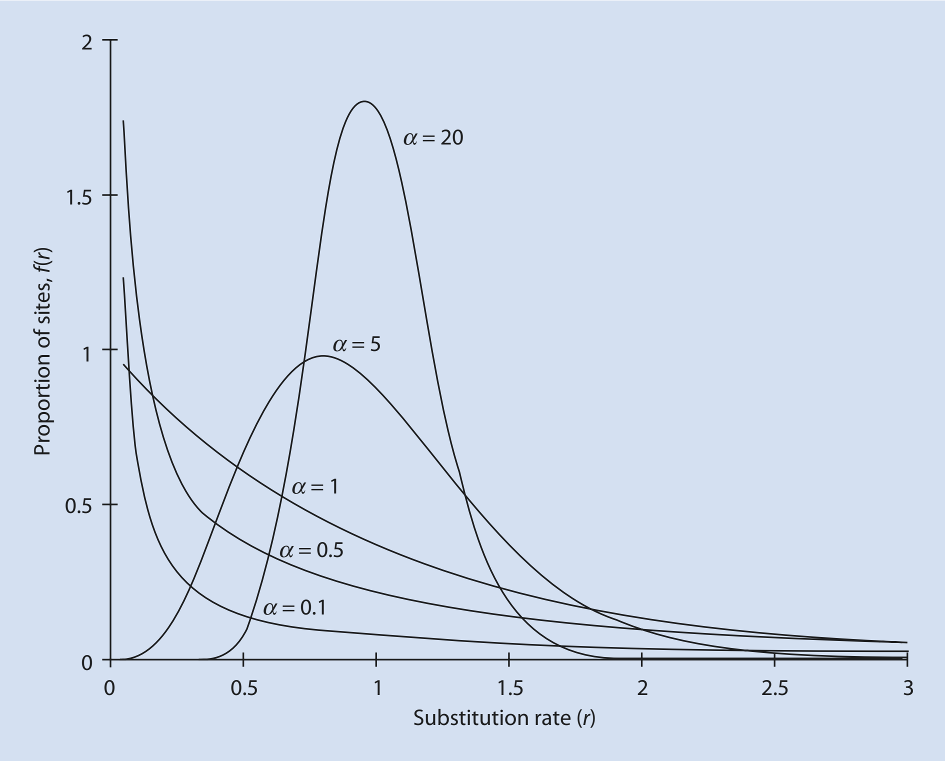


**Fig. 8.6** Probability density function of the gamma distribution. The parameter α defines the shape of the distribution. The *x*-axis shows the substitution rate, whereas the *y*-axis represents the proportion of number of sites with this rate (Reprinted from (Yang 1996) with permission from Elsevier)

sites is already included by the gamma distribution and thereby the estimation of parameters for both model modifications is not independent (Sullivan et al. 1999). However, simulation studies and comparisons of real datasets found that models including both parameters (+ I + Γ) often improve phylogenetic analyses (Jia et al. 2014; Kück et al. 2012).

As outlined, two of the main assumptions of Markov models are stationarity and homogeneity of the data. Moreover, the so far discussed models also assume reversibility, meaning that the probability of a substitution from character i to character j equals the reverse case. However, real data often violates one or several of these assumptions (see above). A model not using these assumptions is the general Markov model (GMM), which has been formalized for phylogenetics by Barry and Hartigan (1987). However, this model is in practice too complex, as it requires the reliable estimation for more parameters than usually available for a given dataset. Two more restricted versions of the GMM were developed by Jayaswal et al. (2011), with one model assuming stationarity, while being nonreversible and non-homogeneous (SBH model), and another model assuming stationarity and reversibility, while being non-homogeneous (RBH model). It has been demonstrated for an example dataset that these models have a better fit to the data than standard GTR models or its derivatives (Jayaswal et al. 2011). However, their use was so far restricted to smaller datasets due to the computational complexity.

## 8.3 Models of Amino Acid Substitutions

Most models for amino acid substitutions can be broadly classified into two classes: empirical models and mechanistic models (Yang 2006). Empirical models are usually derived from a large compilation of sequence alignments. These models are summarized in an amino acid replacement matrix, where each entry is corresponding to the relative rate of replacement of one amino acid by another. The first empirical matrices have been published by Margaret Dayhoff and colleagues (Dayhoff et al. 1972, 1978). The matrices were compiled from available protein alignments of similar sequences which did not differ in more than 15% of their sites. Altogether, 34 protein superfamilies split into 71 alignments have been analysed. For each alignment, a phylogenetic tree using maximum parsimony (▶ see Sect. 8.5) has been created, where internal nodes represent ancestral protein sequences. Mapping all changes on the tree allowed inferring the number of amino acid replacements for all possible pairs. All changes were inferred from sequences with a high identity to reduce the probability of multiple substitutions. As such, all entries of the corresponding matrix are regarded for an evolutionary time interval of 1 amino acid change per 100 amino acid sites (◘ Fig. 8.7). This matrix is known as PAM1 (point accepted mutations) matrix. To derive matrices for sequences separated by a longer time (and experienced more change), the PAM1 matrix can be multiplied by itself. A widely used PAM matrix is the PAM250 matrix (Dayhoff et al. 1978), which has found to be reliable for sequence which differ in up to 80% of their sites.

Jones et al. (1992) published an update of the Dayhoff matrices, based on a much larger database including newly available sequence alignments that fulfilled the original chosen requirements of 85% identity. By using distance methods instead of maximum parsimony, also a slightly different methodology to select pairs of sequences for the final analyses was used. This widely used matrix is known as the JTT model of amino acid substitutions. For both matrices, Dayhoff and JTT, the included phylogenetic analyses methods to count changes along the tree for the generation of the substitution model have been harshly criticized (Whelan and Goldman 2001). As such, it has been noted that both approaches likely underestimate the number of replacements (even for highly similar sequences). Instead, a maximum likelihood (▶ see Sect. 8.5) approach has been proposed, which avoids the outlined problems of the discussed models. This approach is able to use sequences with different degrees of identity and also allows the occurrence of multiple changes (Yang et al. 1998). Using this methodology, Whelan and Goldman (2001) inferred an amino acid replacement matrix based on analysing 182 protein families which is known as the WAG model. A refinement of the WAG matrix using an updated and larger database including nearly 4000 alignments has been published by Le and Gascuel (2008) and is now referred to as the LG model. Later on, Le et al. (2012) introduced the use of different substitution matrices for site with different evolutionary rates. Two sets of matrices called LG4M and LG4X were estimated from a huge number of protein alignments and are used according to different gamma categories (LG4M) or a distribution-free scheme of rate heterogeneity (LG4X). The use of different substitution matrices for different sites (mixture models) has been generally proven to outperform the choice of a single substitution matrix for all sites (Le et al. 2008), but is computationally demanding. All these models are based on sequence data sampled across the tree of life. Several amino acid substitution models have been developed for specific taxa or organelles. These models are solely based on sequence comparisons from the taxon or organelle of interest. Besides others, models are available for mitochondria (MtRev)

| | | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| Ala | A | 9857 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| Arg | R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| Asn | N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| Asp | D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| Cys | C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Gln | Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| Glu | E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| Gly | G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| His | H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| Ile | I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| Leu | L | 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| Lys | K | 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| Met | M | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| Phe | F | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 | 0 |
| Pro | P | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| Ser | S | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| Thr | T | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| Trp | W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Tyr | Y | 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 9945 | 1 |
| Val | V | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9901 |

**Fig. 8.7** Matrix containing probabilities of amino acid substitutions of proteins undergoing 1% of change (1 accepted point mutation in 100 amino acids, PAM1). All values are multiplied by 10,000. For example, there is a 0.02% probability that the amino acid arginine (*ARG*) will be replaced by alanine (*ALA*)

(Adachi and Hasegawa 1996), chloroplasts (CpRev) (Adachi et al. 2000) or retroviruses (RtRev) (Dimmic et al. 2002). Moreover, specific models have been inferred for mammal (MtMam) (Yang et al. 1998) or arthropod mitochondria (MtArt) (Abascal et al. 2007), to just name a few.

The so far described models have fixed substitution rates, corresponding to values in the matrix estimated from a large database. However, it is also possible to estimate these substation rates directly from the data using the GTR model. Given the high number of parameters to be estimated for a 20 × 20 substitution matrix (208 parameters!), this approach makes only sense for huge datasets.

As mentioned for nucleotide substitution models (see above), these models can again be modified according to a gamma distribution (+ Γ) and by including the amount of invariant sites (+ I). Usually, the amino acid frequencies are specified according to the chosen rate matrix. However, it is also possible estimating the specific amino acid frequencies for the analysed protein alignments (marked with «+ F»).

All so far described models assume homogeneity across sites, and the same underlying model of amino acid substitution applies to all sites (but see LG4M and LG4X). In contrast, the so-called CAT models (which have been unfortunately named the same way as the above-mentioned gamma distribution approximation implemented in RAXML) are site-heterogeneous models, which allow modelling of substitution pattern at different sites of a protein alignment by different substitution matrices (Lartillot and Philippe 2004). CAT models use a Dirichlet distribution (Antoniak 1974) to infer the number of different amino acid matrices with different frequencies, as well as the affiliation of each site to a given matrix (class). Different modifications of the CAT model are available. As such, relative amino acid substitution rates for the matrices used during the analyses can be either fixed (CAT model or CAT-F81 model) or estimated from the data during the analysis (CAT-GTR model). A further modification is the CAT-BP model (Blanquart and Lartillot 2008). This model allows the change of substitution models not only across sites (site heterogeneity) but also across lineages (time heterogeneity). This is facilitated by introducing breakpoints (BP) (Blanquart and Lartillot 2006), which allow switching between substitution matrices. In summary, CAT models infer the number of categories of rate heterogeneity and classify all sites of the alignment accordingly, while each category is modelled using its own relative rate matrix of amino acid substitutions. Besides applying it to amino acid data, CAT-GTR models have been furthermore used for analysing nucleotide data.

Mechanistic models include assumptions about biological processes (e.g. sorting amino acids into classes according to their chemical properties) or are formulated at the codon level. Especially codon models have been proven to outperform empiric models (Miyazawa 2013). However, they come with the computational burden of being extremely time-consuming to calculate, as they have to specify a matrix of 61 × 61 possible codon transformations (stop codons are not included) (Zaheri et al. 2014). Widely used is a simplified version of the codon model proposed by Goldman and Yang (1994). In this model, parameters are estimated for codon pair comparisons. A rate of 0 is applied for codons which differ in two or three positions. Separate rates are estimated for codons differing in only one position. In this case, different rates are estimated for synonymous and non-synonymous transversions, as well as for synonymous and non-synonymous transitions (Ren et al. 2005). Moreover, the frequency of codons can be handled differently for this model, either it assumed that all codons have the same frequency (Fequal) or the frequency is estimated based on a set of nucleotide frequencies (F1 × 4), or estimated from three sets of nucleotide frequencies for each codon position (F3 × 4) or estimated directly as codon frequency (F61) (Yang 2006). The names in brackets refer to the number of free parameters to be estimated, which range from 1, over 4 and 12, to 61. Further, different codon models have been proposed (Zaheri et al. 2014). Similar to empirical models for amino acid substitutions, an empirical model of codon substitution has been inferred based on more than 17,000 alignments (Schneider et al. 2005). Kosiol et al. (2007) combined the approach described above with knowledge from empirical models regarding amino acid replacement rates based on physicochemical properties. Finally, Zaheri et al. (2014) published a generalized codon model based on a reduced number of parameters. Whereas most codon models are still computationally too intensive for phylogenomic analyses of large datasets, they are frequently used when detecting adaptive molecular evolution in single genes (Yang and Bielawski 2000), e.g. implemented in the software PAML (Yang 2007).

## 8.4 Model Selection and Data Partitions

### 8.4.1 Model Selection

The selection of the best fitting model for any given dataset is a crucial step in phylogenetics (Anisimova et al. 2013). Nucleotide substitution models differ in the number of parameters estimated from the dataset – and therefore in the way of realistically describing the data. However, there is a trade-off. More parameters allow a more realistic way of representing the underlying data. But this comes with the danger that too many parameters may over-fit the underlying data (overparametrization), resulting in errors during parameter estimation (Sullivan and Joyce 2005). In contrast, simplified models may not realistically represent the data, which can also mislead phylogenetic reconstruction. In the case of amino acids, empirical models mostly differ regarding the database they were compiled from. Moreover, some models have been specially designed for certain taxa or organellar proteins. Besides this, for all models, the question arises if they should account for rate heterogeneity (+ Γ) and invariant sites (+ I), as well as if the frequency of amino acids should be estimated from the data (+ F). Obviously, methods are needed to select a model that fits the underlying data while trying to avoid overparametrization. The most widely used methods to choose among models are the hierarchical likelihood ratio test (hLRT), the Bayesian information criterion (BIC) and the Akaike information criterion (AIC).

A popular approach using hLRT to select the best fitting nucleotide model has been implemented in a software called MODELTEST (Posada and Crandall 1998), which later on also was updated (JMODELTEST2) in including more models and other selection criteria (Darriba et al. 2012). The basic idea behind the hLRT approach is to calculate the likelihood (► see Sect. 8.5) for a fixed topology (e.g. a simple distance tree of the alignment to be evaluated) given the selected model and compare it with the likelihood for an alternative model:

$$\delta = 2\left(\ln L_1 - \ln L_0\right) \tag{8.3}$$

In this formula, $L_1$ represents the more complex model (in terms of free parameters) and $L_0$ the alternative model. The more complex model (which always will result in a better

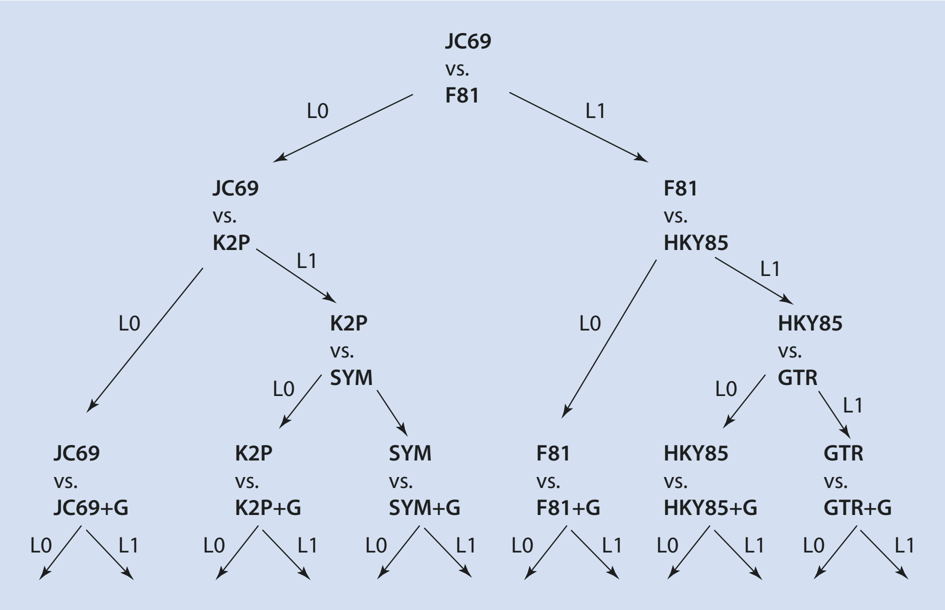


**Fig. 8.8** Hierarchical tree for model testing using hLRT as implemented in MODELTEST (Posada and Crandall 1998). The test starts with the comparison of the two least complex models, and progresses along the tree. If the more complex model is chosen, the *arrow* L1 has to be followed. In case of choosing the less complex model, L0 has to be followed. The full tree includes all models and modifications (+I, +G, +I +G) (not shown), and the model is chosen when testing arrives at the bottom of the tree

likelihood value) will be chosen, if the value $\delta$ is regarded as significant when evaluated by a $\chi^2$ test statistic, where the degree of freedom equals the difference in the number of free model parameters. For example, in the JC69 model, there is one free parameter ($\mu$) to be estimated, whereas in the HKY85 model, there are five free parameters ($\mu$, ti/tv and three base frequency parameters, whereas the frequency of the fourth will be set to add up to 1). Always two models are compared and can be tested along a tree-like hierarchy (Fig. 8.8). Starting with the comparison of the least complex models (JC69 vs. F81), tests are conducted following the tree hierarchy until a model is selected.

By using the hLRT, two models are compared at a time. In contrast, using the information criteria AIC (Akaike 1973) or BIC allows to simultaneously compare all considered models. Moreover, for hLRT, it is important that the models are nested, which means they can be transformed into each other by restricting or opening parameters, as it is the case for nucleotide substitution models. However, amino acid models do not fulfil this criterion. AIC and BIC are able to compare nested and non-nested models. Like hLRT, both information criteria use likelihood scores calculated under the assumption of the model to be tested, which are then penalized according to the open parameters these models use (Posada and Buckley 2004). The AIC is calculated as:

$$\mathrm{AIC} = -2\log_e L_i + 2K_i \tag{8.4}$$

The idea behind the AIC is to test the goodness of fit (represented by the likelihood expressed as $\log_e L_i$ in this formula), by also taking into account the variance of the estimated parameters by the model (given as $K_i$, which represents the number of free parameters estimated by the model). The smaller the AIC, the better is the fit of the model to the data. The BIC is an easy-to-calculate approximation of the Bayes factor (Kass and Wasserman 1995) and is defined as:

$$\mathrm{BIC} = 2\log_e L_i + K_i \log_e n \tag{8.5}$$

In this formula, $L_i$ is the likelihood given the model and the fixed topology, $K_i$ gives the number of free parameters in the model, and n is the length of the alignment (in bp). The BIC measures the relative support of the data for any compared model. Both criteria are implemented in widely used software for the selection of nucleotide (JMODELTEST2) (Darriba et al. 2012) and protein models (PROTTEST3) (Darriba et al. 2011).

### 8.4.2 Partition Finding

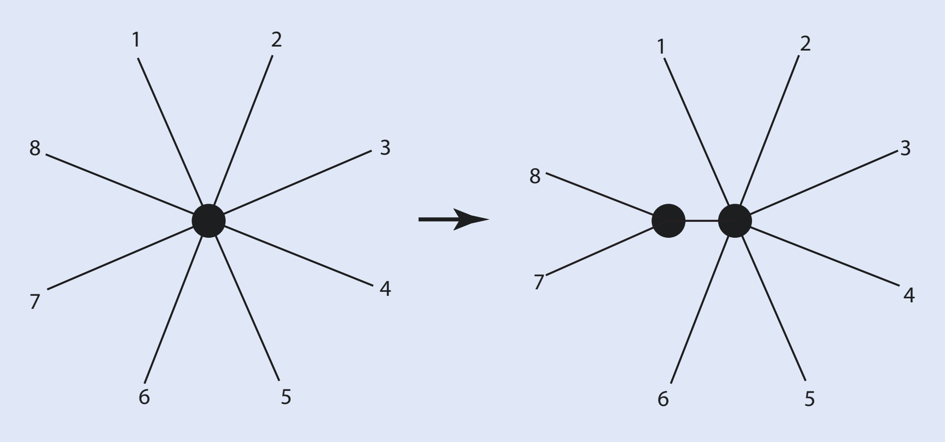
Phylogenomic datasets usually contain hundreds or even thousands of genes (or genetic loci in general). Choosing the same model for the complete dataset is unrealistic, as differences of evolutionary rates across genes or codon positions are to be expected. However, (over-)partitioning by estimating individual models for every gene or locus can easily lead to overparametrization (Li et al. 2008). Consequently, sites or genes that evolve similarly should be merged for model selection. Such an approach is called partitioning, where the dataset is divided into homogenous blocks of sequences which evolve similarly. Consequently, a method for selecting a partition scheme for multigene datasets is needed. Yet, as in the case of possible tree topologies, the number of possible data partitions grows fast with the chosen units. For example, there are >100,000 schemes to partition a dataset of ten genes, and this number grows to more than 100 sextillion possibilities (8,47E + 23) when considering the three different codon positions (30 units) (Li et al. 2008). Lanfear et al. (2012) proposed a heuristic solution to find optimal partition schemes for large datasets, which is computationally manageable and implemented in the software PARTITIONFINDER. As for model testing, a phylogenetic tree is estimated from the data. Given this tree, the best-fit substitution models (as described above) are chosen for the defined units (called subsets, e.g. genes, codon positions). For each subset, the log likelihood (▶ see Sect. 8.5) is estimated. This allows estimating the likelihood of each analysed partitioning scheme by summing up the likelihood scores of the subsets which are part of this scheme. As the number of potential partitioning schemes gets astronomical even for smaller datasets, a heuristic approach using a greedy algorithm is used to limit the number of analysed schemes. Using information criteria like AIC or BIC, the optimal partitioning scheme is chosen. Nevertheless, this approach is still time intensive for large phylogenomic datasets. Consequently, a faster approach suitable for large to very large datasets has been developed based on a hierarchical clustering approach (Lanfear et al. 2014). With this approach, parameters are first estimated for initial data blocks, which are then combined based on their similarity.

## 8.5 Inferring Phylogenies

Four widely used methods for phylogenetic reconstruction will be introduced: neighbour joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). Several other methods (e.g. UPGMA, minimum evolution) have been proposed, but are basically not in use anymore in modern molecular phylogenetics. Inferring phylogenies based on molecular data can be conducted by either using pairwise distances between sequences (NJ) or based on discrete characters (MP, ML, BI) (Page and Holmes 1998). Usually, with distances, the best tree is reconstructed by clustering, whereas character-based methods apply an optimality criterion to choose the best tree(s) among all possible tree topologies (Yang and Rannala 2012). Historically, the first computer-based analyses of sequence data often relied on distances (Fitch and Margoliash 1967a). However, today, character-based methods are clearly favoured for phylogenetic reconstruction.

### 8.5.1 Neighbour Joining

Inferring trees by NJ consists of two steps: construction of a matrix of pairwise distances which is used for a subsequent clustering of a tree using the NJ algorithm, which chooses the tree with the smallest sum of branch lengths (Saitou and Nei 1987). Usually, distances between sequences are calculated by considering an evolutionary model (see above). This matrix is clustered into a tree by the NJ algorithm, which uses star decomposition. The algorithm starts with a completely unresolved star tree and successively joins a pair of terminals based on the distance matrix until the tree is fully resolved (◘ Fig. 8.9). Iteratively, terminals are chosen in a way to minimize the total branch length of the tree. After every step, the distance matrix is updated newly, and the recently joined terminals are also



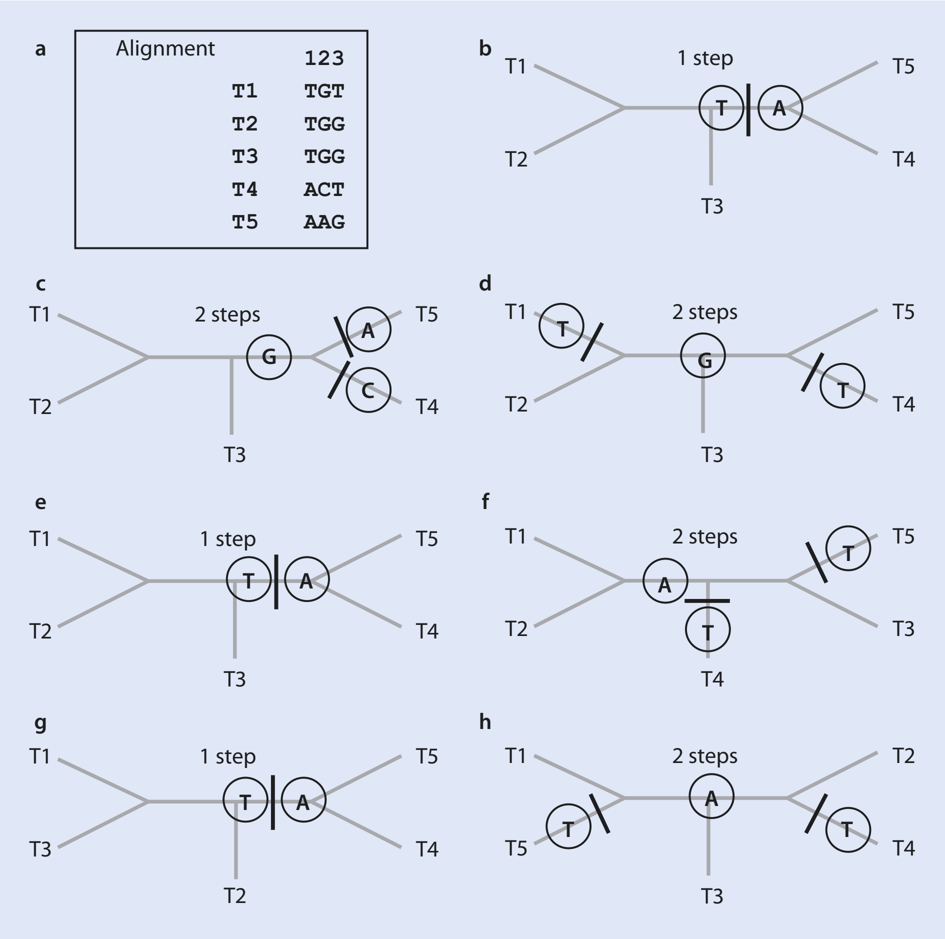
◘ **Fig. 8.9** Star decomposition as conducted by the neighbour-joining algorithm. Based on a distance matrix, the two terminals are joined which maximally reduce the total length of the tree, thereby creating a new internal node. After this step, the distance matrix is updated, and the process is repeated until the tree is completely resolved joined in the matrix as composite terminals. A detailed description of the algorithm is given in Nei and Kumar (2000).

The NJ algorithm is, for example, implemented in the software MEGA7 (Kumar et al. 2016) or PAUP* (Swofford 2003). NJ is computationally superfast, as the time for analysing large datasets can still be measured in (mili)seconds. However, distance methods in general have been shown to be prone to problems with systematic errors and missing data (Brinkmann et al. 2005) and are therefore rarely used for phylogenomic analyses. Nevertheless, this method is often implemented when a quick tree is needed, e.g. guide trees for alignments or starting trees for heuristic searches of character-based methods (see below).

### 8.5.2 Maximum Parsimony

MP is a phylogenetic inference method using an optimality criterion to decide which trees are the best among all possible trees. As the number of possible trees for larger numbers of analysed sequences is too big to be analysed exhaustively, heuristic methods are used to narrow the space of searched trees (see below). The explicit rational behind MP is the idea that the best hypothesis to explain an observation is the one which requires the fewest assumptions (Steel and Penny 2000). This rational goes back to the medieval Franciscan friar William of Ockham («Ockham's razor») and is now widely used as a scientific method in general. For molecular phylogenetics, MP as method for reconstructing trees was basically introduced by Edwards and Cavalli-Sforza (1963), even though they called it minimum evolution (not to be confused with the distance-based minimum evolution method proposed by Rzhetsky and Nei (1992)!). A couple of years later, Camin and Sokal (1965) also published a parsimony-based reconstruction method, as well as Fortran-based computer programs called CLADON I to III, to carry out the steps of the analysis. Nowadays, there are several different variants of MP in use, which, for example, differ in the way if character transformations are weighted or ordered (Felsenstein 1983). In the following, the so-called Fitch parsimony is explained, where a change between any two character states is possible and all changes count equally (Fitch 1971). For MP analysis, the character states for every single alignment site (character) are mapped on a tree topology while minimizing and counting the assumed changes (steps) (◘ Fig. 8.10). For example, in ◘ Fig. 8.10b–d, the different characters are mapped onto the same topology, and the number of transformations (steps) is counted. This MP score is measured across all possible topologies, and the trees with the lowest number of steps are chosen as the most parsimonious trees. Only characters that produce different numbers of steps across topologies are regarded as informative (e.g. ◘ Fig. 8.10e–h), whereas all other characters are excluded from the analysis. Informative characters are those which have at least two different character states, which appear at least in two terminals each. The most widely used programs for MP analyses are PAUP* (Swofford 2003) and TNT (Goloboff et al. 2008).

MP is a method which is easy to understand, and due its simplicity, efficient and fast algorithms for analysis are available (Yang and Rannala 2012). However, the lack of an explicit use of evolutionary models is a major drawback for this method. Comparisons of model-based (e.g. ML) and MP inference have been extensively discussed in the literature and especially the journals *Cladistics* and *Systematic Biology* represented a battleground for proponents of these methods in the late 1990s and early 2000s. Most simulation studies show that model-based approaches based on ML inferences (including BI)



Alignment

|  | 123 |
|---|---|
| T1 | TGT |
| T2 | TGG |
| T3 | TGG |
| T4 | ACT |
| T5 | AAG |

**Fig. 8.10** Most parsimonious reconstructions of character change measured in steps. **a** Example alignment. **b**–**d**. Character transformations for alignment positions 1–3 reconstructed on the same unrooted tree. **e**–**h**. Reconstruction of the same alignment position on different unrooted tree topologies, illustrating that the same character can produce different numbers of steps

outperform MP in molecular phylogenetic reconstruction (Felsenstein 2013; Huelsenbeck 1995). However, MP methods are widely used for the phylogenetic reconstruction of absence/presence patterns of genome level characters, e.g. retrotransposons or microRNAs.

### 8.5.3 Maximum Likelihood

The likelihood function is defined as the probability of the data given the underlying parameters and was originally developed by the statistician R. A. Fisher in the 1920s. In a phylogenetic context, a tree topology represents a model, whereas the branch lengths of this topology and the underlying substitution parameters are parameters of this model (Yang and Rannala 2012). In an ML analyses, the tree topology and its set of branch lengths

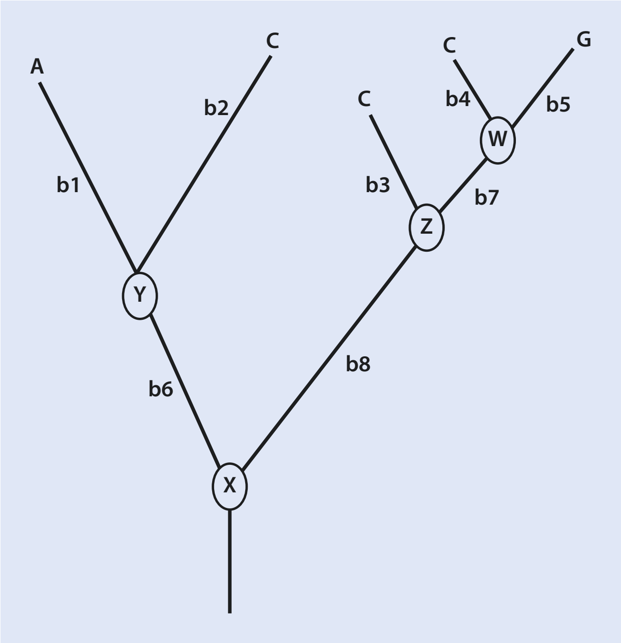


**Fig. 8.11** Computing the likelihood for a single alignment site on a fixed tree using Felsenstein's pruning algorithm (see text). The letters at the tips represent character states of the terminals for this topology. Letters in circles represent ancestral nodes; b1–b7 denote branches and its corresponding lengths

are searched for, for which the data (the sequence alignment) most likely evolved as we observe it. As such, ML analyses comprise two steps. First, for a given tree topology, the lengths of individual branches as well as the parameters for an evolutionary model of sequence evolution have to be optimized. The latter part is usually conducted during the model testing procedure, as described above. Second, the most likely topology across all possible topologies has to be found using the likelihood *L* as an optimality criterion. The calculation of *L* is very time-consuming; however, Felsenstein (1981) has introduced a ML algorithm (pruning algorithm) for molecular phylogenetics, which has made ML analyses feasible. Using this approach, it is assumed that the evolution at different sites and across lineages is independent. First, the likelihood for a single site for a given topology, given branch lengths (b1–b8 in Fig. 8.11) and a chosen evolutionary model, is calculated (Fig. 8.11). The probability for a single site is the sum of the probabilities of each scenario, overall possible nucleotides that may have existed at the interior nodes (w, x, y, z in Fig. 8.11). This means, computing from the tips to the root, the probability for the presence of every possible character state for each internode has to be calculated, given the underlying substitution model. The algorithm is explained in detail in several textbooks (Felsenstein 2013; Nei and Kumar 2000; Yang 2006). The likelihood *L* for a given tree for the complete alignment is the product of the site-wise likelihood calculations. As these numbers are very small, usually the negative logarithm of the likelihood is used. The topology which produced the best likelihood value is finally chosen by the optimality criterion.

ML analyses are the state of the art for phylogenomics, and most publications in this field use this approach. In the early 2000s, using ML was still often computationally difficult. However, with improvements of computer technology, availability of high-performance computing clusters (HPC cluster) and especially software, which leverages this development, ML analyses became feasible for even very large datasets. At the forefront of developing user-friendly software that can also be run on HPC clusters is Alexandros Stamatakis, the developer of the software RAXML (Stamatakis 2014). This program has been well adapted to the environment of HPC clusters, and a related software (EXAML) has been published for phylogenomic analyses on supercomputers (Kozlov et al. 2015). Both programs come with the caveat that for nucleotide analyses only the GTR model (and modifications) can be chosen. Alternative programs for large-scale ML analyses include PHYML (Guindon et al. 2010), FASTTREE (Price et al. 2010) and IQ-TREE (Nguyen et al. 2015). The latter program has also a user-friendly way of finding partitions and models for large datasets implemented.

### 8.5.4 Heuristic Methods and Genetic Algorithms

Computing likelihoods and optimizing branch lengths for a tree topology is time-consuming. Conducting these operations for all possible topologies to choose the tree that has the highest likelihood is basically impossible for even smaller datasets. Similarly, for MP analyses, it is impossible to investigate all possible topologies for larger datasets. For this reason, heuristic methods which only investigate a fraction of all trees, while at the same time enhancing the chance that this fraction contains the best tree, are used. Typically, a reasonable starting tree is computed to begin the heuristic search. For example, in the widely used software RAXML (Stamatakis 2014), this starting tree is inferred using a MP analysis, but it can also be based on NJ or chosen randomly. Most heuristic methods use rearrangement operations to change this starting tree and to compute new trees for phylogeny inference. Using specific rearrangement rules, different but always feasible numbers are generated based on the starting tree. The most popular heuristic tree rearrangement operations are nearest neighbour interchange (NNI), subtree pruning and regrafting (SPR) and tree-bisection and reconnection (TBR) (Felsenstein 2013). NNI swaps adjacent branches of a tree. Using SPR, a subtree of the tree is removed and regrafted into all possible positions. By TBR, a tree is split into two parts at an interior branch, and all possible connections between branches of these two trees are made. NNI will produce the smallest number of rearranged trees and TBR the largest number. After performing the possible rearrangements, the tree with the best likelihood value is chosen. Using this tree, a new round of rearrangements is performed, and the process is repeated until no better trees are found. Several modified versions of these basic operations exist. All these methods try to limit the tree space in a way that the best tree is still found. However, as there is no guarantee to find the best tree using heuristics, it is strongly recommended to conduct several replicates of the phylogenetic analysis to enhance the chance of finding the best solution.

Alternative ways for heuristic searches of ML analyses are genetic (or evolutionary) algorithms (GA). By using GA, trees represent individuals within a population, whereas the likelihood function is used as a proxy for the fitness of each individual. Fitter trees will produce more offspring trees, which are allowed to mutate over generations (e.g. by using rearrangement operations). A selection step randomly choses rearranged and unchanged trees which will be kept in the next generation. The evolution of trees is monitored over many generations, and after stopping this procedure, the tree with the highest likelihood is chosen. A GA algorithm for ML search is, for example, implemented in METAPIGA (Helaers and Milinkovitch 2010).

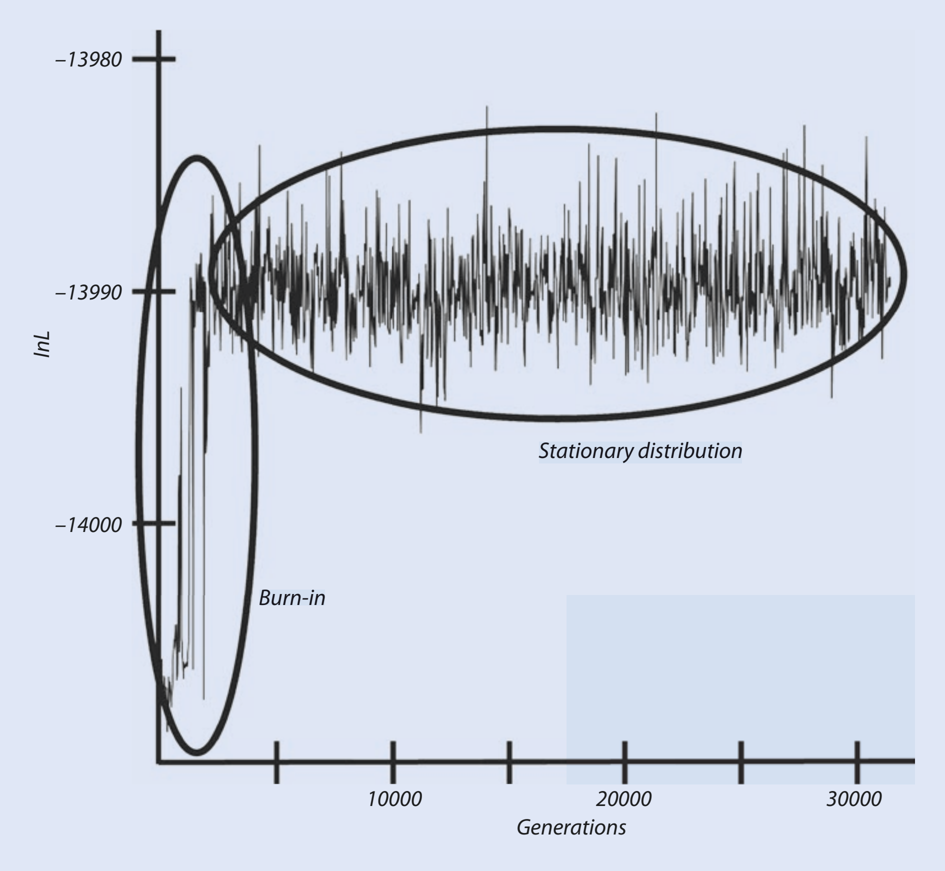
### 8.5.5 Bayesian Inference

Whereas the likelihood describes the probability of observing the data given a hypothesis (and evolutionary model), by using BI, the probability of the hypothesis given the data is described. For BI, prior probabilities and posterior probabilities have to be distinguished. Prior probabilities are assumptions made before the BI analyses. These prior probabilities are then updated according to the analysed data, and posterior probabilities are the result of BI. Using Bayes' theorem (Formula 8.6) in a phylogenetic context, the posterior probability ($f(\theta|X)$ can be calculated by multiplying the prior probability for a tree (and its parameters) ($p(\theta)$) with the likelihood of the observed data (given a tree and its parameters) ($l(X|\theta)$, and as a denominator, a normalizing constant of this product is used ($\int p(\theta)$ $l(X|\theta)d\theta$) (Yang and Rannala 2012).

$$f(\theta\,|\,X)=\frac{p(\theta)l(X\,|\,\theta)}{\int p(\theta)l(X\,|\,\theta)d\theta} \tag{8.6}$$

Obviously, by using this approach in phylogenetics, different assignments of prior probabilities to tree hypotheses would have a huge impact on the posterior probabilities. To circumvent this problem, so-called flat priors are used, where all tree topologies have the same prior probability (Huelsenbeck et al. 2002b). Accordingly, all differences in the posterior probability can be attributed to differences in the likelihood function. However, there is a profound difference how both analyses use parameters of the models of sequence evolution. ML conducts a joint estimation, where the likelihood for all parameters is optimized at once. In this case, the likelihood of one parameter is dependent on the likelihood estimation of every other parameter. In contrast, BI uses a marginal estimation, where the posterior probability of any one parameter is calculated independently of any other parameter. So, even by using flat priors and identical models, ML and BI might infer different phylogenetic trees due to the differences between joint and marginal likelihood estimation (Holder and Lewis 2003).

Solving Bayes' theorem analytically is computationally too intensive. However, an approximation of posterior probabilities by using a Markov chain Monte Carlo (MCMC) approach made BI of phylogenies feasible (Larget and Simon 1999). By using a Markov chain, a series of random variables is generated, and the probability distribution of future states is only dependent on the current state at any point in the chain. For inference of phylogenies, the Markov chain starts with a randomly generated tree including branch lengths. The next step in the chain is to generate a new tree, which is based on the previous tree (e.g. using tree rearrangement heuristics or changing branch length parameters). This is called a proposal. The proposed new tree is accepted given a specific probability based on the Metropolis-Hastings algorithm (Holder and Lewis 2003). Roughly spoken, this means that it will be usually accepted if it exhibits a better likelihood and only sometimes accepted, when it has a worse likelihood. If the proposed tree is accepted, it will become the new current state to propose the next step in the chain. If the newly proposed tree is rejected, the current tree remains, and a new tree has to be proposed for the next step. Running such a Markov chain will quickly generate better trees. However, under specific conditions, there will be no better trees found, and Markov chain will have a «stationary distribution». At this point, all trees (topologies plus branch lengths) sampled are expected



■ **Fig. 8.12** Likelihood scores of a MCMC run plotted against generations. Once stationarity is achieved, trees from this distribution are sampled by discarding all other trees as burn-in. A majority-rule consensus of sampled trees will provide posterior probabilities for every node

to be close to the optimum, and the number of how often a tree has been visited by the chain is interpreted as an approximation of the posterior probability of this tree. By sampling a number of trees from this stationary distribution (all other sampled trees are discarded as *burn-in*) (■ Fig. 8.12), a majority-rule consensus tree can be generated, where the frequency of each node approximates its posterior probability (Huelsenbeck et al. 2002b). As this approach might be problematic if the chain runs into local optima, usually four Markov chains (and two independent analyses) are run in parallel (Metropolis-coupling, MCMCMC). These chains are differently explorative regarding the tree space (hot chains), and only one chain is used for sampling trees to infer the posterior probability distribution (cold chain). However, the chains are in contact and are allowed to swap their status every *n* generations (Huelsenbeck and Ronquist 2001). A widely used software for BI of phylogenies is MRBAYES (Ronquist et al. 2012). With REVBAYES, a major rewrite of this program has been published (Höhna et al. 2016). Moreover, the program PHYLOBAYES uses BI and has the site-heterogeneous CAT model of sequence evolution integrated (see above) (Lartillot et al. 2009, 2013). The program BEAST uses BI to generate ultrametric trees for molecular clock analyses (Drummond et al. 2012).

Bayesian analyses of phylogenomic datasets have been especially used for molecular clock analyses, where the program BEAST (Drummond et al. 2012) became widely popular. For standard analyses with the aim of retrieving a tree topology with support values (see below), ML seems to be the better alternative, as it is computationally usually faster, whereas the results are often similar. The biggest problem of BI is the question how long chains have to run to become stationary. Several metrics have been published to diagnose stationarity (Nylander et al. 2008), but for large datasets, this becomes difficult. Furthermore, posterior probabilities seem to overestimate the node support (Alfaro et al. 2003; Erixon et al. 2003; Simmons et al. 2004), which usually makes it necessary to either way run a ML analysis with bootstrapping (► see Sect. 8.6).

## 8.6 Support Measures

Phylogenetic analyses will always result into a tree topology, which raises the major question how much trust can be put into it. Usually, the most interesting is the support for certain interior branches of the tree. One measure, posterior probabilities, has been already introduced in ► Sect. 8.5. The most common measure of support for phylogenies is derived from bootstrap analyses. The bootstrap is a resampling technique commonly used in statistics for estimating the variability of an estimate (Efron 1982). The application of bootstrapping for phylogenetics was introduced by Felsenstein (1985). To conduct bootstrap analyses the original dataset has to be resampled with replacement. As such, so-called pseudoreplicates are generated which contain the same number of alignment sites as the original alignment. Every site in these pseudoreplicates is filled by sites from the original alignment. As this sampling is conducted with replacement, the pseudoreplicates may include some original sites multiple times, where others could be missing. Normally 100 or 1000 pseudoreplicates are generated, which are then analysed as in the original phylogenetic analysis (e.g., with ML, MP or NJ). Alternatively, a «bootstopping» algorithm can estimate the number of necessary replicates (Pattengale et al. 2009). All trees resulting from these analyses are then summarized as a majority-rule consensus tree and the frequencies are given at the nodes. If a branch is found in all replicates, the support is 100%. In statistics, these values are interpreted in the typical fashion that values >95% are statistically significant. This support describes how well a branch is supported by the data, not the probability if a branch is «true». This also implies that the bootstrap basically tests if the dataset is large enough to recover a well-supported solution. Earlier studies dealing with single gene datasets claimed that the bootstrap might be over-conservative, and values >70% can be regarded as significant (Hillis and Bull 1993). However, large datasets as used in phylogenomics seem to inflate highly supported branches, and a 95% threshold of support seems reasonable here. Nevertheless, it should be kept in mind that especially systematic error within the data can lead to significantly supported branches which are wrong. Bootstrap analyses are computationally time intensive, and several approaches which are able to quickly approximate bootstrap values for large datasets have been published, e.g. implemented in IQ-TREE (Minh et al. 2013) and RAXML (Stamatakis et al. 2008). A related resampling method that has been used in phylogenetics is the jackknife, where instead of resampling with replacement, randomly half of the positions are deleted in the pseudoreplicates (Felsenstein 2013).

An alternative way of estimating branch support is based on likelihood ratio tests (► see also Sect. 8.4). In the case of the approximate LRT (aLRT), the idea is based on comparing internal branches of an inferred tree to the null hypothesis, where the length of this branch is zero (Anisimova and Gascuel 2006). However, for testing purposes, the null hypothesis of a zero branch length is approximated by testing against the putatively incorrect branching. For this, the best topology is compared with the best alternative arrangement around the branch of interest. For any given internal branch, only three topological arrangements are possible in its neighbourhood, which can be easily ordered by their likelihood. The LRT test statistic is calculated as two times the difference in likelihood between the best tree (*L1*) and the best alternative hypothesis (*L2*). The result is compared against a mixed $\chi^2$-distribution. Simulation studies show that the aLRT is much faster and similarly accurate as standard bootstrap approaches, as long as the underlying evolutionary model of the phylogenetic analysis is not strongly violated (Anisimova et al. 2011). A possibility to get a more robust version of the LRT under the presence of model misspecifications is the inclusion of a bootstrapping step in this test. This has been done for the SH-aLRT (Guindon et al. 2010), where a variant of the bootstrap called RELL is used (Kishino et al. 1990). RELL (resampling estimated log likelihoods) is a shortcut to calculate likelihood values for pseudoreplicates. Instead of generating pseudoreplicates of the alignments, site-wise likelihoods of the best tree of the original alignment are bootstrapped. This fast (but maybe inaccurate) method helps to generate a distribution of likelihoods for a large number of pseudoreplicates, without having to perform the time-consuming ML optimization step. The SH-aLRT compares the distribution of the RELL-bootstrapped topologies with a test statistic developed by Shimodaira and Hasegawa (1999). Simulation studies have shown that the SH-aLRT is much more robust for datasets analysed under model violations than the aLRT (Anisimova et al. 2011). LRT for branch support is, for example, implemented in PHYML (Guindon et al. 2010) and IQ-TREE (Nguyen et al. 2015).

## 8.7 Molecular Clocks

According to the molecular clock hypothesis, which assumes a constant rate of evolution over time, it is possible to date divergence times in phylogenetic trees using the fossil record (Hasegawa et al. 1985). The existence of a molecular evolutionary clock was first hypothesized by Zuckerkandl and Pauling (1965), based on the results of their landmark study which proposed the existence of a uniform rate of evolution among globin genes in different species (Zuckerkandl and Pauling 1962). This result was in line with the finding of Doolittle and Blomback (1964), who found an inverse relationship of species divergence time and difference in protein sequences. However, with the availability of more DNA sequence data, it became obvious that mutation rates can be different across taxa and genes, thereby implying that a strict molecular clock hypothesis is an unrealistic assumption (Kumar 2005). Several methods have been developed to deal with this problem. Sarich and Wilson (1973) and Fitch (1976) proposed a relative-rate test, where the rate of evolution of two (ingroup) sequences is independently compared to an outgroup sequence. By this procedure, it is possible to test if the distance between the two ingroup sequences to its last common ancestor is equal (or not significantly different), as assumed under the molecular clock hypothesis. If a $\chi^2$-test indicates a significant difference in this distance, it is rejected that this pair of sequences evolves according to a molecular clock. With the help of such a test, gene alignments (and included sequences) can be filtered, and only those who fulfil the molecular clock criterion are used for analysis. Relative-rate tests demonstrated that the assumption of a global molecular clock is unrealistic for most datasets, thereby prohibiting molecular clock analyses for them. However, local molecular clock analyses within a maximum likelihood framework are possible, where some lineages evolve under different evolutionary rates, while other lineages have a constant rate (Yoder and Yang 2000). Sanderson (1997) developed a method (nonparametric rate smoothing), which is based on the assumption that evolutionary rates show autocorrelation over time. This idea goes back to Gillespie (1991), who suggested that substitution rates evolve among lineages and are inherited from ancestors to descendants. Under this assumption, a model can be used which minimizes the change of evolutionary rate between related (ancestor-descendant) lineages while allowing variation across lineages. Nowadays, most widely used are Bayesian approaches which allow the use and comparison of alternative models of substitution changes over time and for different data partitions (Lepage et al. 2007), as, for example, implemented in the software BEAST (Drummond et al. 2012).

An obviously important step for every molecular clock analysis is the calibration of the resulting ultrametric tree. This is usually done by using fossil data, but also biogeographic events can be helpful. Till the end of the 1990s, it was a commonplace to use a single calibration point for molecular clock analyses. Often, a single gene was analysed with the help of one dated internal node, where the substitution rate of the dated lineage was divided by the age of the dated divergence to subsequently transform all genetic distances into absolute time (Renner 2005). Later on, it became standard to use multiple calibration points (if available!), which could be used to cross validate each other (Benton et al. 2009). Moreover, it is possible to assign minimum and maximum ages for any used calibration point. Minimum ages are hard bound, indicating that a certain clade must have at least this age as evidenced by the first appearance in the fossil record. In contrast, maximum ages are more difficult to assign and are thereby soft bound, given as a distribution, which tries to estimate the origin of a species which is always certainly older than its first appearance in the fossil record (Donoghue and Benton 2007). A best practice guide for the justification of fossil calibration has been published by Parham et al. (2012).

The potential and pitfalls of molecular clock analyses are nicely illustrated by several studies dealing with the origin of animals. It always has been a conundrum that animal fossils are either rare or disputed (e.g. the Ediacaran fauna) in the Precambrian (>541 mya) fossil record, whereas basically all major phyla are suddenly found in different Cambrian (541–485 mya) Lagerstaetten (Briggs 2015). This conundrum is known as the «Cambrian explosion». Molecular clocks represent an interesting approach to investigate the timing of animal evolution, and many publications dealing with this topic have been published in the last decades. However, instead of converging to a similar conclusion, most of these studies differ wildly in their results. As such, dates for the emergence of animals range from ~600 mya (Peterson et al. 2004) to ~1300 mya (Hedges et al. 2004). Moreover, often these dates come with a huge error rate, making precise statements difficult (Graur and Martin 2004). These errors are often introduced due to the problems of assigning well-supported calibration dates for such old fossils, questioning the possibility of using molecular clocks for rejecting or supporting hypothesis of early animal evolution in general (dos Reis et al. 2015). However, many examples of dating younger divergences clearly emphasize the power of molecular clock analyses, which have been used for less controversial divergence time estimates of the evolution of, e.g. insects, mammals, humans or plants (dos Reis et al. 2016; Renner 2005). Moreover, molecular clock analyses have been successfully used to analyse virus outbreaks, as in the case of Ebola, HIV or influenza (dos Reis et al. 2016).

## References

Abascal F, Posada D, Zardoya R (2007) MtArt: a new model of amino acid replacement for arthropoda. Mol Biol Evol 24:1–5

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348–358

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd international symposium on Information Theory. Budapest, p 267–281

Alfaro ME, Zoller S, Lutzoni F (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol 20:255–266

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55:539–552

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60:685–699

Anisimova M, Liberles DA, Philippe H, Provan J, Pupko T, von Haeseler A (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. BMC Evol Biol 13:161

Antoniak C (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174

Barry D, Hartigan JA (1987) Statistical analysis of hominoid molecular evolution. Stat Sci 2:191–207

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10

Benton MJ, Donoghue PCJ, Asher R (2009) Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S (eds) The timetree of life. Oxford University Press, Oxford, pp 35–86

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Blanquart S, Lartillot N (2006) A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol 23:2058–2071

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842–858

Briggs DEG (2015) The cambrian explosion. Curr Biol 25:R864–R868

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–757

Camin JH, Sokal RR (1965) A method for deducing branching sequences in phylogeny. Evolution 19:311–326

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772–772

Dayhoff M, Eck R, Park C (1972) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC, pp 89–99

Dayhoff M, Schwarz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55:65–73

Donoghue PCJ, Benton MJ (2007) Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends Ecol Evol 22:424–431

Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci U S A 104:2043–2049

Doolittle RF, Blomback B (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. Nature 202:147–152

dos Reis M, Thawornwattana Y, Angelis K, Telford Maximilian J, Donoghue Philip CJ, Yang Z (2015) Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol 25:2939–2950 dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17:71–80

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Edwards AWF, Cavalli-Sforza LL (1963) The reconstruction of evolution. Heredity 18:553

Efron B (1982) The jackknife, the bootstrap and other resampling plans. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia

Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol 52:665–673

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

Felsenstein J (1983) Parsimony in systematics: biological and statistical issues. Annu Rev Ecol Evol Syst 14:313–333

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

Felsenstein J (2013) Inferring phylogenies. Sinauer Associates, Sunderland

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

Fitch WM (1976) Molecular evolutionary clocks. In: Ayala FJ (ed) Molecular evolution. Sinauer Associates, Sunderland, pp 160–178

Fitch WM, Margoliash E (1967a) Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279–284

Fitch WM, Margoliash E (1967b) A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65–71

Fourment M, Gibbs MJ (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol Biol 6:1

Fryxell KJ, Moon W-J (2005) CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22:650–658

Gillespie J (1991) The causes of molecular evolution. Oxford University Press, New York

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

Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24:774–786

Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet 20:80–86

Gu X, Fu YX, Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546–557

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

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

Heads M (2005) Dating nodes on molecular phylogenies: a critique of molecular biogeography. Cladistics 21:62–78

Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2

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

Hess PN, De Moraes Russo CA (2007) An empirical test of the midpoint rooting method. Biol J Linn Soc 92:669–674

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756–766

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F (2016) RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 65:726–736

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275–284

Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Huelsenbeck JP, Bollback JP, Levine AM (2002a) Inferring the root of a phylogenetic tree. Syst Biol 51:32–43

Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002b) Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51:673–688

Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python environment for tree exploration. BMC Bioinformatics 11:24

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

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R (2007) Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460

Huson DH, Rupp R, Scornavacca C (2010) Phylogenetic networks. Concepts, algorithms and applications. Cambridge University Press, Cambridge

Jayaswal V, Jermiin LS, Poladian L, Robinson J (2011) Two stationary nonhomogeneous markov models of nucleotide sequence evolution. Syst Biol 60:74–86

Jia F, Lo N, Ho SYW (2014) The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 9:e95722

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci CABIOS 8:275–282

Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro R (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479

Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31:2577–2579

Krell F-T, Cranston PS (2004) Which side of the tree is more basal? Syst Entomol 29:279–281

Kück P, Mayer C, Wägele J-W, Misof B (2012) Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLoS One 7:e36593

Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6:654–662

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:82

Larget B, Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–759

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

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288

Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611–615

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320

Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Philos Trans R Soc Lond Ser B Biol Sci 363:3965–3976

Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936

Lepage T, Bryant D, Philippe H, Lartillot N (2007) A general comparison of relaxed molecular clock models. Mol Biol Evol 24:2669–2680

Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245

Li C, Lu G, Ortí G (2008) Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst Biol 57:519–539

Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195

Miyazawa S (2013) Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic analysis. BMC Evol Biol 13:257

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

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 32:268–274

Nixon KC, Carpenter JM (1993) On outgroups. Cladistics 9:413–426

Nylander JAA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583

Page RDM (1996) Tree view: an application to display phylogenetic trees on personal computers. Compu Appli Biosci CABIOS 12:357–358

Page RD, Holmes E (1998) Molecular evolution: a phylogenetic approach. Blackwell, Osney Mead/Oxford

Parham JF, Donoghue PCJ, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, Patané JSL, Smith ND, Tarver JE, van Tuinen M, Yang Z, Angielczyk KD, Greenwood JM, Hipsley CA, Jacobs L, Makovicky PJ, Müller J, Smith KT, Theodor JM, Warnock RCM (2012) Best practices for justifying fossil calibrations. Syst Biol 61(2):346–359

Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A (2009) How many bootstrap replicates are necessary? In: Batzoglou S (ed) RECOMB 2009, LNCS 5541. Springer, Berlin/Heidelberg, pp 184–200

Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA (2004) Estimating metazoan divergence times with a molecular clock. Proc Natl Acad Sci U S A 101:6536–6541

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16:37–45

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490

Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1:53–58

Ren F, Tanaka H, Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol 54:808–818

Renner SS (2005) Relaxed molecular clocks for dating historical plant dispersal events. Trends Plant Sci 10:550–558

Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142:485–501

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

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945

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

Sanderson M (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14:1218

Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144–1147

Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6:134

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

Simmons MP, Pickett KM, Miya M (2004) How meaningful are Bayesian support values? Mol Biol Evol 21:188–199

Stamatakis A (2006) Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Proceedings of the 20th IEEE international parallel & distributed processing symposium (IPDPS2006). IEEE Computer Society Press, Washington, pp 278–286

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57:758–771

Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol 17:839–850

Sullivan J, Joyce P (2005) Model selection in phylogenetics. Annu Rev Ecol Syst 36:445–466

Sullivan J, Swofford D, Naylor G (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol Biol Evol 16:1347

Swofford D (2003) PAUP*: phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Sunderland

Tavare S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures Math Life Sci (Amer Math Soc) 17:57–86

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

Wilkinson M (1994) Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. Syst Biol 43:343–368

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

Yang Z (2006) Computational molecular evolution. Oxford series in ecology and evolution. Oxford University Press, Oxford

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303–314

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600–1611

Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17:1081–1090

Zaheri M, Dib L, Salamin N (2014) A generalized mechanistic codon model. Mol Biol Evol 31:2528–2541

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genetic heterogeneity. In: Kasaha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New Yoek, pp 189–225

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence, in proteins. In: Bryson V, Vogel H (eds) Evolving genes and proteins. Academic Press, New Yoork, pp 441–465