



# Using models of nucleotide evolution to build phylogenetic trees

David H. Bos<sup>a,\*</sup>, David Posada<sup>b</sup>

<sup>a</sup>*School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, New Zealand*

<sup>b</sup>*Departamento de Bioquímica, Genética e Immunología, Facultad de Ciencias, Universidad de Vigo, Vigo 36200, Spain*

Received 10 February 2004; revised 17 June 2004; accepted 31 July 2004

Available online 21 September 2004

## Abstract

Molecular phylogenetics and its applications are popular and useful tools for making comparative investigations in genetics; however, estimating phylogenetic trees is not always straightforward. Some phylogenetic estimators use an explicit model of nucleotide evolution to estimate evolutionary parameters such as branch lengths and tree topology. There are many models to choose from, and use of the optimal model for a particular data set is important to avoid a loss of power and accuracy in phylogenetic estimations. Here, we review some molecular evolutionary forces and the parameters included in some common models of evolution used to interpret resulting patterns of molecular variation. We present some statistical methods of selecting a particular model of nucleotide evolution, and provide an empirical example of model selection. Statistical model selection strikes a balance between the bias introduced by some models and the increased variance of parameter estimates that results from using other models.

© 2004 Elsevier Ltd. All rights reserved.

**Keywords:** Bayesian phylogenetics; Nucleotide substitution models; Model selection; Akaike information criterion; Likelihood ratio test; Molecular evolution; LMP7

## 1. Introduction

The use of molecular phylogenetics has become widespread in immunological research because

phylogenetic trees are an intuitive way to infer relationships among copies of a gene or among loci of a multigene family. Historically, the primary interest in constructing trees was the pattern of evolutionary relationships itself, or simply the topology of the tree. More recently however, phylogenetic trees are being generated to derive information regarding the processes responsible for the observed pattern of evolutionary relationships, and the tree topology becomes the framework upon which further inference can be drawn. As such, phylogenetics facilitates analysis of gene duplications, rates of evolution, polymorphisms, recombination, divergence of lineages and population

*Abbreviations:* AIC, Akaike information criterion; BIC, Bayesian information criterion; hLRT, hierarchical likelihood ratio test; *I*, proportion of invariable sites; ln *L*, log likelihood; LRT, likelihood ratio test; ML, maximum likelihood; MP, maximum parsimony; NJ, neighbor joining; *ti*, transition; *tv*, transversion.

\* Corresponding author. Address: Department of Forestry and Natural Resources, Purdue University, 715 W. State St, West Lafayette, IN 47907-2061, USA. Tel.: +1 765 494 9779; fax: +1 765 496 9461.

*E-mail address:* [dbos@purdue.edu](mailto:dbos@purdue.edu) (D.H. Bos).

**Symbols**

$\alpha$       alpha shape parameter  
 $\Gamma$       gamma distribution

$\delta$       difference of values  
 $\chi^2$       Chi square distribution

demographics [1,2]. Accurate estimates of evolutionary parameters often hinge on the validity of a single phylogenetic reconstruction upon which inference is based. Inaccurate estimation of trees may lead to biased results and erroneous inference of processes or mechanism of evolution.

Several methods of estimating phylogenetic trees are available. Some of the more commonly used methods include neighbor joining (NJ) [3], maximum parsimony (MP) [4] and maximum likelihood (ML) [5]. More recently, new methods that employ a Bayesian statistical approach [6,7] have been successfully implemented, and these methods have quickly generated much interest [2,8]. While several differences exist, one common feature that unites NJ, ML and Bayesian methods is the use of explicit statistical models of nucleotide evolution.

In the context of phylogenetics, a model provides a framework through which the phylogenetic construction method estimates parameters used to find the preferred tree. The model represents the footprint of evolutionary phenomena that has generated the observed sequence data, such as mutation, selection, and genetic drift. The particular model selected for a data set depends on features of the data such as the level of variation and nucleotide frequencies. While it is not our intent to engage in a full review of phylogenetic methods (for reviews see [9–11]), ML, NJ and Bayesian methods generally benefit from their use of models of evolution in terms of flexibility and performance [12,13].

At the outset, the reconstruction of molecular phylogenetic relationships seems a relatively simple exercise. However, the intricacies of DNA sequence evolution and the culmination of molecular forces acting on sequences can make phylogenetic inference a complex matter. The purpose of this paper is to highlight the uses and advantages of nucleotide models in light of the complexities of evolutionary genetics. First we review aspects of DNA sequence evolution such as rates of evolution and changes in

those rates through time and along the sequence. We then examine parameters of some models commonly used in phylogenetics that correspond to aspects of sequence evolution and discuss model selection and use. Finally, we present an empirical example of model selection in comparative immunology and use it to demonstrate how results can vary depending on the model being used and argue that appropriate model selection and use is critical to accurate scientific exploration of genetic information.

## 2. Sequence evolution and phylogenetics

### 2.1. Substitutions

As more DNA sequences become available, it is apparent that patterns of nucleotide changes used to construct trees are very complex. These complexities arise because of a number of factors contributing to and acting on the primary unit of sequence differences—substitutions. Substitutions can be classified as transitions (ti) or transversions (tv) (Fig. 1). Transitions are substitutions between structurally similar nucleotides (e.g.  $A \leftrightarrow G$ , which are both purines), and transversions occur between dissimilar

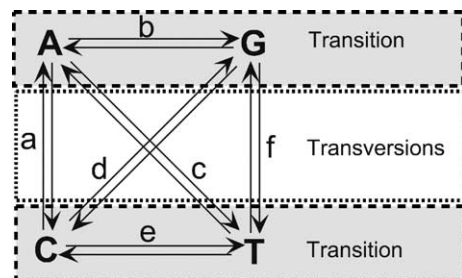


Fig. 1. A substitution matrix representing the possible different rates of evolution for the two possible transitions, and four possible transversions (a–f). In this substitution matrix, substitution parameters are reversible, so that the rate of change from nucleotide  $i$  to nucleotide  $j$  is the same as rate of change from  $j$  to  $i$ .

nucleotides (e.g. A ↔ T; purine to pyrimidine). Transitions are often observed at more than two times the rate of transversions (ti: tv > 2) even though there are twice as many possible transversions for any given nucleotide site. This trend towards more transitions occurs because mutation to a similar nucleotide is more likely to be tolerated than a dissimilar one, and this transition bias can be quite pronounced in some molecules, especially mitochondrial DNA. Frequently, whether or not a substitution is a transversion has implications for altering the protein coded by a DNA sequence.

Substitution rates can vary along a DNA sequence in at least two different ways. First, because of the redundant nature of the genetic code, substitutions are similarly tolerated more or less in various positions within each codon [14]. For instance, the third position of a codon evolves much more quickly than the second position because substitutions at the second position usually change the amino acid encoded by that codon, while similar substitutions at the third position do not. Second, to preserve the function of the protein, its structure must be conserved in important regions; other segments of the protein may be less conserved (for a well known example, see [15]). Thus substitution rates vary in different parts of the DNA sequence correlating to different domains in the protein (i.e. among codons rather than within a codon) and can cause different parts of a gene to support different trees. The variation in substitution rates among different nucleotides in a sequence (rather than in a codon) is referred to as substitution rate heterogeneity or among-site rate variation. In a DNA sequence with among-site rate variation, some nucleotide sites undergo frequent substitutions, while others may change very slowly or not at all [16]. The occurrence of among-site rate variation alters the probabilities of nucleotide substitutions from the often-assumed notion that substitutions are randomly spread along the sequence, and is nearly ubiquitous among DNA sequences [17,18].

## 2.2. *The molecular clock*

The idea of the molecular clock is based on early observations that the number of amino acid replacements between species or lineages is proportional to the divergence time between them [19]. The empirical

observation of a molecular clock was explained by the neutral theory of molecular evolution [20], where such a clock would be expected if most amino acid substitutions were selectively neutral, driven by mutations and random drift. Although the neutral theory has become pervasive in evolutionary genetics, the molecular clock does not always tick regularly [21]. Variation of substitution rates both within a lineage and among lineages makes the existence of a global molecular clock unlikely even though neutral mutations may dominate molecular evolution. Anything that changes the balance between drift and selection can alter the tick-rate of the molecular clock by causing a temporary increase or decrease in the number of substitutions per unit of time, and even neutral evolution can occur in an episodic manner [22,23]. Events such as gene or genome duplications, speciation or changes in the population size can change the dynamic between drift and natural selection, altering the rate of evolution if only for a short period of time.

Many lines of evidence are against a universal molecular clock; however, neutral theory still plays a prominent role in evolutionary genetics. The action of natural selection does not imply that neutral substitutions do not exist, only that they do not always accumulate with clock-like regularity. Violations of the molecular clock are commonly found in highly divergent gene sequences, genes that are the product of gene duplications [24], or genes that have experienced natural selection or changes in structure or function [25]. There are many difficulties associated with using a molecular clock [26], nevertheless, it is often the case that tests of the molecular clock [27] cannot reject clock-like evolution for closely related gene sequences. This could indicate that molecular evolution is clock-like for periods of evolutionary time, or that methods may lack statistical power to reject a molecular clock in some cases. Even when clock-like evolution is plausible, precise estimation of dates can still be difficult to obtain because of different assumptions and sources of uncertainty [28]. Also, methods are available that relax the assumption of a strict molecular clock and allow one to estimate evolutionary dates in lineages that have different rates [29–31].

Many evolutionary processes create irregular patterns of nucleotide substitution and the detection

and characterization of these irregularities has led to a better understanding of DNA sequence evolution. In turn, our understanding of molecular evolutionary patterns has allowed us to develop statistical models used to represent the irregularities of DNA sequence evolution. For instance, through the use of these models, researchers are able to overcome common phylogenetic scenarios that are positively misleading for methods that do not use statistical models such as MP [32–34]. Although models are ultimately major simplifications, summarizing many evolutionary forces and events, appropriately incorporating these models generally leads to improvement of genetic distance and phylogenetic analysis [11].

### 3. Models of nucleotide substitution

#### 3.1. Phylogenetic estimators

Neighbor Joining, ML and Bayesian methods all rely on explicit statistical models of evolution to reconstruct evolutionary trees. The Neighbor Joining algorithm is different from ML and Bayesian methods because it uses the model to calculate pairwise genetic distances between sequences, and reconstructs a topology based on those distances. Maximum likelihood and Bayesian methods use the sequence data directly to reconstruct a tree, thereby utilizing information in specific nucleotide differences instead of summarizing changes with a genetic distance. Due to these differences, ML offers noteworthy statistical properties in comparison with genetic distance-based methods, but is much more computationally intensive [32,35,36]. While NJ and ML methods are well understood and their uses are common in the literature, Bayesian methods are relatively new.

The Bayesian method is related to ML method because they both utilize the likelihood function. However, when using Bayesian statistics to reconstruct a phylogeny, the preferred outcome is the one that maximizes the posterior probability, which is determined by the prior distribution and the likelihood of that tree. The prior distribution for trees, models and parameters can be specified to be generally uninformative to avoid bias, or it can reflect prior knowledge from other sources. Whereas other

methods produce a single best estimate of evolutionary relationships and ignore uncertainty of the final outcome, Bayesian methods produce a set of trees of which the one with the highest posterior probability is accepted as the preferred tree. Bayesian methods are generally faster than ML methods, and also offer the advantage of automatically incorporating an estimate of phylogenetic uncertainty [6]. While many aspects of Bayesian phylogenetic estimation have yet to be refined and explored, these methods offer the same benefits from employing statistical models as ML and NJ [6,7]. These benefits include the flexibility to incorporate a wide range of models, easy hypothesis testing, and improvements on estimates of numbers of substitutions, efficiency and robustness [37].

#### 3.2. Model parameters

Statistical models of nucleotide change represent aspects of the pattern of variation that results from the process of evolution. Models vary in complexity according to the number of parameters used to represent evolutionary change. While simple models summarize nucleotide substitutions with one or two parameters, the most general models can involve more than 60 parameters (e.g. codon models that are introduced below). Model parameters can reflect differences in nucleotide frequencies, substitution rate (such as transition bias) and among-site rate variation. The substitution matrix of a model represents different rates of evolution between certain pairs of nucleotides, and the gamma distribution models among-site rate variation. In other words, the substitution matrix determines the substitution rate between specific nucleotide pairs (e.g.  $A \leftrightarrow G$ ), and the gamma distribution determines the overall substitution rate at a nucleotide site. Combining different parameters has resulted in a large number of models, but many of them share several parameters (Table 1).

The JC69 model [38] considers all possible nucleotide substitutions to have an equal probability, and is the simplest available model (Table 1). Felsenstein [5] suggested a model in which probabilities of nucleotide changes were determined by the equilibrium nucleotide frequencies. Kimura [39] proposed a model that utilizes a relatively simple substitution matrix that allows for two different rates: one for transitions and the other for transversions.

Table 1  
Some commonly used nucleotide models and summary of parameters

Model	Parameters			
	Number of parameters	Nucleotide frequencies	Substitution rate in Fig. 1	Reference
JC69	1	Not included	$a=b=c=d=e=f$	[38]
F81	4	$\pi_A, \pi_C, \pi_G, \pi_T$	Not included	[5]
K80	2	Not included	$a=c=d=f, b=e$	[39]
K81	3	Not included	$a=f, b=e, c=d$	[40]
HKY85	6	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b=e$	[42]
SYM	6	Not included	$a, b, c, d, e, f$	[108]
TrN	7	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b, e$	[44]
GTR	10	$\pi_A, \pi_C, \pi_G, \pi_T$	$a, b, c, d, e, f$	[109]

Parameters of these models can include four different base frequencies and up to six substitution rates. Flexibility of models is such that invariable sites and/or a gamma distribution can simply be added to incorporate rate variation.

Kimura [40] and others [41,108] have also formulated models that incorporate more than two rates in the substitution matrix, thus enabling models to account for different rates of change between all of the possible nucleotide pairs. In an effort to make models more representative of empirical observations, Hasegawa et al. [42], Felsenstein [43], Tamura and Nei [44] and Rodriguez et al. [109] each created models which incorporate multiple aspects of sequence evolution (Table 1). These models combine parameters for differences in substitution rates and differences in nucleotide frequency.

Among-site rate variation can also be incorporated into models of nucleotide evolution. The simplest way to statistically represent among-site rate variation is to divide sites into two classes: those that vary and those that are invariable. To better account for wide rate differences among the variable sites, several methods have been used [44,45], but the most successful involves the use of a gamma distribution [18,46]. The gamma distribution can be approximated with as little as four categories [47], and the statistical representation of rate variation is independent of substitution models like those described above and can simply be added to any pre-existing model (for example, we can specify a JC69 +  $\Gamma$  model).

Under the gamma distribution, there is a continuum of probabilities of change for nucleotides, ranging from low to high. The numbers of nucleotide sites with the various rates of substitutions determines the shape of the gamma distribution that is summarized by the shape parameter ( $\alpha$ ). When most of the nucleotides are invariable or have very slow rates,

then the shape of the distribution is skewed to the right (Fig. 2). Under this scenario there are a few nucleotides with high rates and the shape parameter would be small ( $\alpha < 1$ ), indicating a high level of rate variation, i.e. not all nucleotides evolve at a similar rate. As a result, most of the variation in the data set comes from relatively few nucleotide sites that are evolving very rapidly (substitutional ‘hotspots’).

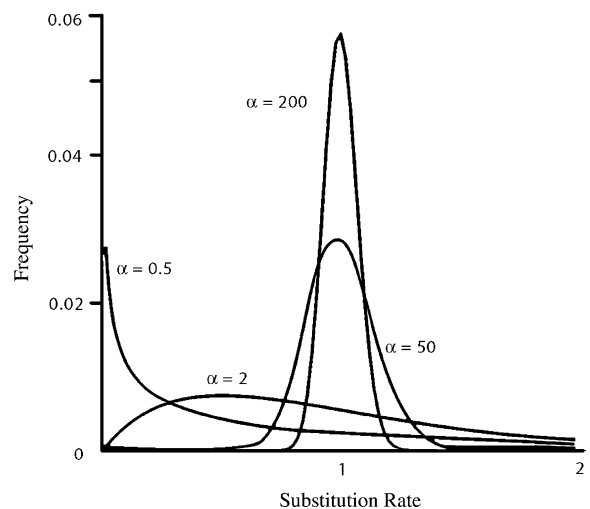


Fig. 2. Gamma distributions calculated using different shape parameters ( $\alpha$ ). The number of nucleotides in a sequence evolving at a particular rate determines the shape parameter. When a sequence contains mostly invariable nucleotide sites and variation is concentrated at a few rapidly evolving nucleotide sites, the shape parameter is small ( $< 1$ ). As the proportion of variable nucleotide sites increases, the shape parameter becomes larger, indicating that more sites evolve at a moderate rate and fewer sites have extremely high or low rates.

Large shape parameters ( $\alpha > 20$ ) indicate a more bell-shaped distribution with most sites having intermediate rates of evolution with few nucleotides evolving at very high or low rates (Fig. 2). As the shape parameter becomes larger, more nucleotide sites have a more similar rate of evolution and among-site rate variation becomes increasingly inconsequential [13].

The above-mentioned model parameters all work at the individual nucleotide level, and therefore treat each nucleotide as an independent unit. However, for protein coding DNA sequence this is not the case. Whether or not a substitution changes an amino acid depends on the other nucleotides in that codon when the substitution occurs, thus individual nucleotide sites in protein coding sequence are not independent. To accommodate this, nucleotide models that treat a codon triplet as an independent unit have been formulated to more accurately model coding DNA [48–50]. Variations of these models provide parameters to account for transition bias, codon frequency, rate variation among codon positions, and different rates for non-synonymous substitutions [51]. Codon models can become very complex by parameterizing each codon frequency, but these models can also approximate codon frequencies with fewer parameters. Unfortunately, these models are generally not implemented for use in reconstructing phylogenetic trees except when using some Bayesian methods [7]. Instead, codon models have been typically used to estimate substitution rates and detect levels of natural selection acting on a protein.

### 3.3. *Effects of models*

The performance of a model-based phylogenetic method may depend on the fit of the model to the data [10]. Similarly, the efficiency of distance-based methods is dependant on the accuracy of model-based estimates of genetic distance [11]. For sets of sequences that are long with low levels of polymorphism, the model may have little effect on the outcome of analysis. However, when working with more divergent sequences, the use of one model over another can alter the results of analysis, and even lead to strong support for the wrong tree topology [52], a fact that underscores the importance of using the best-fit model for a particular data set. Due to the wide diversity in size, variation and rates of evolution

among different data sets, there is no single best-fit model suited for use in any data set. Use of inadequate, overly simplistic models selected without statistical validation often leads to biased estimation of evolutionary genetic parameters [12,33,37,53,54].

The model parameter with one of the strongest influences on genetic distance and phylogenetic estimation is among-site rate variation. Rate variation among sites is particularly problematic and misleading when substitution rates also vary among branches in the tree (e.g. non-clock-like evolution) [32]. When both types of variation are present, use of the best fit model seems to be essential to obtain the correct tree topology [16,55]. Except in cases with strong rate variation among both sites and lineages, tree topology estimation is relatively robust to violations of model assumptions [36,56]. Unfortunately the same robustness does not extend to estimation of parameters such as substitution rates, branch lengths and genetic distance. Failing to include rate heterogeneity among sites results in underestimation of the number of substitutions at highly mutable sites [16]. Consequently, branch lengths are underestimated, and this effect is much more prominent in longer branches than shorter ones [54]. This is likely to be due to the fact that phylogenetic estimators give greater weight to highly variable sites in a sequence [47].

Simplifying the assumptions of a model by failing to include a factor for transition bias can also adversely alter the outcome of analysis. A transition bias is found universally among DNA sequences [57] and inclusion of this parameter is essential for accurate estimates of genetic distance for NJ analysis [58,59]. Similarly, failure to incorporate transition bias will result in underestimation of branch lengths in ML phylogeny estimation [60]. Aside from the inherent problems of branch length and genetic distance underestimation, these factors can alter the tree topology and lead to erroneous conclusions regarding the dates of lineage splitting [44]. There is also an interplay between transition bias and among-site rate variation, so that the level of among-site rate variation is underestimated (overestimation of  $\alpha$ ) using models that exclude a transition bias [60].

One of the major advantages of using models is the ability to more accurately estimate the actual number of substitutions that have occurred in a set of sequences. This allows researchers to include

sequences of high variability because homoplasy in the form of superimposed substitutions can be accounted for with the use of models. The alternative way of dealing with sites or sequences which are suspected of saturation of substitutions, is simply to eliminate them from consideration. While this does effectively eliminate the influence of homoplasy at those sites, any information that can be gleaned from those sites is also lost and the size of the sample is decreased, exposing the analysis to the increasing effects of sampling error or bias.

While potential problems with simple models are documented, some also dispute the utility of more general models [61]. Some criticisms of very complex models point out that these models have greater difficulty distinguishing between tree topologies because of smaller differences in likelihood scores, and that as more model parameters are added, more error is associated with each parameter estimate. These properties of complex models are general statistical phenomena and are not limited to phylogenetic analysis; however, while these points are valid, they arise because of random rather than systematic error. As a result these problems can be mediated rather than aggravated by addition of data [13]. The amount of data required for consistent phylogenetic analysis depends on the shape of the tree, numbers of taxa and levels of diversity. If the tree shape is not symmetric and branch lengths are very long, then analysis of data with less than 500 nucleotides will generally not be reliable, especially for more general models [33,62]. Consistency and reliability of phylogenetic inference is expected to increase by analyzing longer sequences and additional taxonomic sampling.

The potential bias introduced through using a particular model also has an effect upon the level of support given to a tree topology with techniques like bootstrapping [63]. The most widely accepted interpretation of the bootstrap is that it is the level of support for a particular node of a tree that the data provides [64]. As such, it represents whether the same topology might be recovered if more data are collected, rather than if the relationship is correct. However one interprets the bootstrap values, the accuracy and precision of the bootstrap values depends on the fit of model [65]. For instance, if a phylogenetic method or a model is used that has systematic bias, then the bootstrap will also reflect

that bias [13]. Consequently, bootstrap values used in such a case will be artificially high and reflect strong support for incorrect branching patterns.

Bayesian methods estimate a level of phylogenetic support that is seen as an intuitive measure of uncertainty regarding each tree topology. Less work has been done to evaluate Bayesian measures of support and the relationship of model specificities and levels of support [66]. However, some research shows that Bayesian measures of support are good estimates of phylogenetic accuracy [67], but others conclude that these values are overestimates of the true level of uncertainty [68,69]. Regardless of the procedure used to measure phylogenetic support, caution interpreting results is warranted and use of a statistically rigorous method of selecting a model is recommended.

Although conflicting examples of model complexity and phylogenetic accuracy can be found [65,70], one trend that has emerged is that because of the increase in variance, very short sequences (which are statistically equated with small sample sizes) often do not support the use of the same level of model complexity as longer sequences. Even though the underlying evolution of short sequences may be just as complex as longer sequences, the larger variance inherent with generalized models and small sample sizes makes these types of data more prone to the effects of over-parameterization [71]. While the relationship of model parameters and performance of Bayesian, ML and NJ tree estimation is not always straightforward, a trade-off between the bias of simple models and the increased variance of more general models is generally observed [12]. Consideration of models should take into account the size of the data set, level of divergence, amount of differences in substitutions between different nucleotides, and constancy of rate of evolution both in time and along the sequences. Use of objective criteria to select models will help avoid problems associated with model over-fitting by ensuring that models are not excessively complex and avoid phylogenetic bias by selecting more realistic models [72].

Models that can be implemented in popular phylogenetics programs such as PAUP\* [73], PHYLIP [43], MEGA2 [74] and MRBAYES [7,75] are useful approximations of DNA sequence evolution. Use of one particular model versus another often changes the outcome of analysis, and the choice of

models can be more important than the method of phylogenetic reconstruction. Given that the model plays a great role in the results of analysis, it seems that the choice of one model over another should be justified in some way. Unfortunately, it is still commonplace for models to be used indiscriminately and without justification. The question then becomes, which model is appropriate for a particular data set and how can that model be justified?

### 3.4. Model selection and use

To minimize adverse effects of model over-fitting and model under-fitting, the ideal use of models is to incorporate as much model complexity as needed and no more. Fortunately, methods for selecting the most appropriate model for a particular data set have been proposed. These methods provide a rigorous statistical framework in which to select and justify the best fit model. With the goal of finding the simplest model that accurately approximates sequence evolution, Rzhetsky and Nei [76] developed statistics for selecting models. These tests are independent of evolutionary time and do not require an a priori phylogeny on which to base inference. While this method is computationally efficient, its application is

model-specific and restricted to a limited subset of the available models.

Another method is to use the likelihood ratio test (LRT) to compare models [10]. The LRT statistic is calculated by obtaining the likelihood scores of a null model ( $L_0$ ) and an alternative model ( $L_1$ ). The two scores are then compared by taking twice the difference in the logarithm of the likelihoods to obtain the statistic [ $\delta = 2(\ln L_1 - \ln L_0)$ ]. Use of the LRT in phylogenetics is commonplace for hypothesis testing and the distributions and performance of the test have been investigated [77,78]. When the models compared are nested (one is a special case of the other), the Chi-square distribution ( $\chi^2$ ) is a good approximation of the null distribution of the LRT statistic (df = the difference in the number of free parameters in the two models). In some special cases, fixing one of the parameters of the more parameter-rich model at either boundary (0 or  $\infty$ ) reduces the model to the simpler null model, and a mixed distribution is used [79].

The LRT can be performed on any of the available models, but it requires an a priori input phylogeny to estimate the likelihood of the models [80]. It is also easy to test several models against each other in a series of LRTs that can be performed in a hierarchical fashion (Fig. 3). The likelihood scores of the two

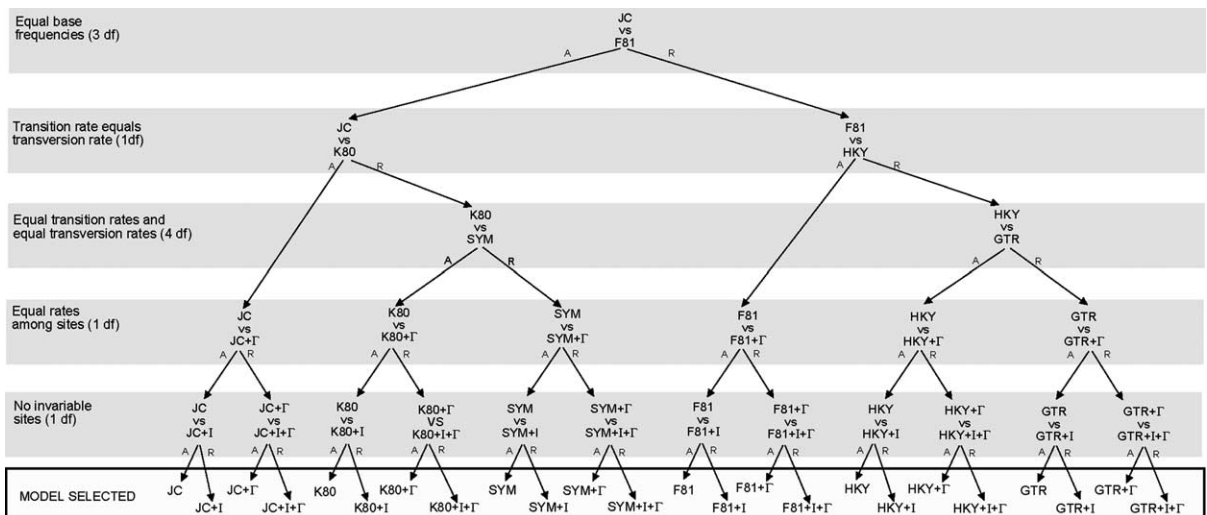


Fig. 3. A ‘decision tree’ of a hierarchical likelihood ratio test. Hypotheses tested are indicated on the left, and this schematic begins with the simplest model and progresses to more complex models in a stepwise manner. The pathway chosen depends on acceptance or rejection of LRT scores, based on a chi squared distribution ( $P < 0.01$ ). In order to preserve the clarity of the figure, not all available models are shown. Models depicted here are: JC [38]; F81 [5]; K80 [39]; HKY [42]; SYM [108]; GTR [109].  $I$ , proportion of invariable sites;  $\Gamma$ , gamma distribution of rates among sites.

Table 2

Several models are compared successively to determine the best fitting model for a data set, starting with the simplest model and increasing complexity

Null model	Alternative model	Parameter tested	LRT ( $\delta$ )	<i>P</i> -value
JC69	F81	Equal base frequencies	4.680	0.196
JC69	K80	ti = tv	90.022	0.000
K80	SYM	Equal ti and tv rates	50.340	0.000
SYM	SYM + $\Gamma$	Equal rates among sites	314.512	0.000
SYM + $\Gamma$	SYM + $\Gamma$ + I	No invariable sites	14.866	0.000

The parameter being tested is assumed by the current null model but not the alternative model. The null model is rejected when the *P*-value of the LRT is  $<0.01$  using a  $\chi^2$  or mixed  $\chi^2$  distribution.

models are compared using the LRT test statistic,  $\delta$ , and significance of the LRT statistic is determined. The better fitting model is retained, it becomes the null model, and the process is iterated with successively more general models of evolution until the addition of further complexity in the alternative model does not create a significantly better fit to the data (Fig. 3 and Table 2). The LRT may be appealing, but the significance of LRTs are easily calculated only for nested models and the a priori distribution of significance for non-nested comparisons is not well established. Performance tests of the LRT also show that this criterion is good at recovering the model used to simulate the sequence data [80], although we should keep in mind that in reality the true model of nucleotide substitution is unknown, and it is much more complex than any candidate model that we can select.

Another way of selecting the most appropriate model for a data set is to use the Akaike information criterion (AIC) [81], which can be thought of as the amount of information lost when a particular model is used to approximate reality. The AIC implements best-fit model selection by calculating the likelihood of proposed models, and imposing a penalty based on the number of model parameters. Parameter-rich models incur a larger penalty than more simple models so that fitting an excessively complex model is not likely. The best fitting model is the one with the smallest AIC value, ( $AIC = -2 \ln L_i + 2N_i$ ), where  $L_i$  is the likelihood for model  $i$  and  $N_i$  is the number of free parameters in model  $i$ . Although the use of LRTs is much more extended in phylogenetics than the use of the AIC, the latter offers important advantages [71]. The AIC is able to compare non-nested models and

simultaneously compares all candidate models, rather than performing sequential pair-wise comparisons; the AIC also has a simple adjustment that more heavily penalizes complex models for data comprised of small samples (i.e. short sequences). The AIC also allows for model selection uncertainty and model averaging. In addition, the AIC recognizes that the true model is not among the set of candidate models so it tries to find the candidate model that best ‘approximates’ the true unknown model of molecular evolution given the amount of information in the data. The objective of model selection is to find the model that will allow one to most accurately estimate unknown phylogenetic parameters while avoiding bias and excessive variance. The model that is best suited to that end will not be an exact representation of cumulative evolutionary processes, but a useful approximation that is appropriate for the level of polymorphism and size of the data set.

Bayesian statistics have also been adapted for use in phylogenetic model selection. Bayes Factors make pairwise model comparisons and are therefore analogous to the LRT procedure [82,83]. Alternatively, the Bayesian information criterion (BIC) can be used [84]. This method more easily enables comparisons of multiple models and is easy to calculate. The posterior probabilities of Bayesian statistics are already used to discriminate between phylogenetic trees and these measures can also be used to choose among multiple models [85]. Like the AIC, Bayesian methods allow estimation of model uncertainty and allow estimation of a phylogeny using a set of candidate models in a model averaging procedure. An important distinction of Bayesian statistics is that calculation of likelihoods proceeds

differently, so that likelihood values compared using Bayesian methods are different from those used in AIC or LRT comparisons.

The above techniques compare model fitness relative to other candidate models, but measuring overall adequacy of a model can also be done. To do this, Navidi et al. [86] and Goldman [87] describe a test that compares a model with an unconstrained model and the appropriate distribution to test significance. Also, Bayesian methods have recently been adapted to examine the adequacy of models [88]. While the unconstrained model is very complex, it is worth noting that when comparing any two models, only aspects in which the models differ are tested. Any aspects models have in common or aspects that are not included in either model remain untested. The outcome of general adequacy tests may find that the selected model is not a complete representation of the data. This is usually thought to be the result of the stringency of the test, instead of gross misrepresentation of the data by the model. Rather, this outcome simply means that the model does not perfectly describe all of the underlying processes of molecular evolution, as would be expected.

The impact of models on phylogenetic analysis is very significant, strongly affecting branch lengths and often topology as well. The use of any particular model is not wrong per se, but we advocate statistical, objective selection among available candidate models to maximize the use of available models for each data set. Unfortunately the model used for analysis is often not justified or even reported in the literature despite its influence on the outcome. However, easy-to-use computer programs that implement rigorous statistical selection of models are available [89]. In the following we demonstrate their use and show how model selection determines the outcome of phylogenetic analysis.

## 4. Empirical example

### 4.1. Data

To illustrate aspects of model selection, we reconstructed the phylogenetic relationships of nine taxa using DNA sequences of the LMP7 gene

downloaded from the Genbank database (accession numbers: human, *Homo sapiens* BC001114 (the human LMP7 is also termed PSMB8 or RING10); mouse, *Mus musculus* U22032; frog, *Xenopus laevis* D44540; salmon, *Salmo salar* AF184938; zebrafish, *Danio rerio* AF032390; medaka, *Oryzias latipes* D89725; pufferfish, *Fugu rubripes* AJ271723; nurse shark, *Ginglymostoma cirratum* D64057; horn shark, *Heterodontus francisci* AF363583). Copies of the gene are from a variety of vertebrates from which full length cDNA was obtained, and the estimated phylogenetic relationships could be used in the framework of studying multigene family evolution, estimating substitution rates, or establishing homology of gene copies. The leader peptide was excluded from analysis, leaving only the coding sequence from the mature protein. The sequences were aligned using Clustal W [90] and alignments were inspected to ensure that the integrity of the coding frame was preserved. The best-fit model for these data was selected using the LRT and AIC after calculating likelihood scores of 24 models using PAUP\*4.0 [73].

### 4.2. Methods

The best fitting model for these data was evaluated according to a hierarchical LRT. The AIC method of model selection was also used to find the best-fit model by calculating the likelihood and subsequently the AIC score of all models. Phylogenetic trees were also calculated using 24 models of evolution selected to represent a variety of statistical complexity. These models have an arbitrary relationship to the data, and the resulting trees can be compared to those obtained using models selected using rigorous statistical criteria. Here, we use the ML method of phylogenetic construction as implemented in PAUP\* [73] because it is known to be robust to violations of model assumptions and because the statistics of ML estimation are well understood [10,36,56]. We calculated these scores manually to demonstrate the method, but the program MODELTEST [89] provides the appropriate command block for PAUP\* to automatically calculate the likelihood scores for 56 models, which can then be automatically compared using the LRT and AIC in the MODELTEST program.

### 4.3. Results

The model selected by both the LRT and AIC is the SYM model with both invariable sites and a gamma distribution of among-site rate variation (SYM+ $\Gamma$ +I; see Tables 2 and 3) [47,108]. This model includes a substitution matrix allowing for six different rates of substitutions: one for each type of reversible nucleotide change. There is no significant heterogeneity of nucleotide frequencies accounted for in the model, but the model makes provisions for considerable rate variation along the gene sequence (see Table 3). The invariable sites of the sequence alignment are accounted for in the model and the gamma distribution represents rate heterogeneity only among variable sites. As the distribution of the gamma shape parameter is skewed towards the right, most of the variable nucleotides evolve fairly slowly, with a few sites evolving more rapidly. Models with few parameters commonly used to reconstruct phylogenetic relationships were rejected by both selecting criteria in favor of more general models (Table 2).

A test of the overall adequacy of the preferred model against the unconstrained model [87] indicates a sufficient level of support for the SYM+ $\Gamma$ +I model. The test statistic of the difference in likelihoods between the unconstrained and SYM+ $\Gamma$ +I models was 1091.546. Monte Carlo simulations under the null (SYM+ $\Gamma$ +I) model hypothesis were done to determine the null distribution of differences in

likelihood between the unconstrained and SYM+ $\Gamma$ +I models. This distribution ranged from 946.723 to 1196.620 with a mean value of 1069.492. The test statistic falls well within the 95th percentile of the distribution, indicating that the null hypothesis (SYM+ $\Gamma$ +I) cannot be rejected against the unconstrained model under these criteria. The best-fit model selected above was therefore used to reconstruct the phylogenetic relationships among these taxa, and the result indicates a topology consistent with generally accepted relationships (Fig. 4).

When the evolutionary relationships among these genes were estimated using other models, three different tree topologies emerged (models and likelihood scores found in Table 4). Many simple models rejected by statistical model selection criteria preferred a tree in which the frog and shark share a most recent common ancestor, and this clade is a sister group to a clade in which mammals and teleost fish

Table 3  
Molecular evolutionary parameter values of best-fit model, SYM+ $\Gamma$ +I, selected under the LRT and AIC criteria

Parameter	Value
<i>Substitution matrix</i>	
A: C	1.49
A: G	2.12
A: T	1.53
C: G	0.62
C: T	4.24
G: T	1.00
<i>Base frequencies</i>	
A	0.25
C	0.25
G	0.25
T	0.25
Proportion invariable sites	0.411
Gamma shape parameter	2.827

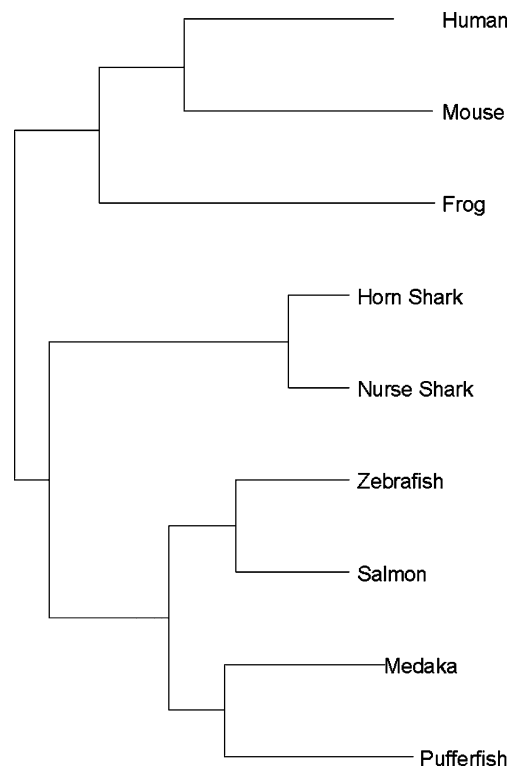


Fig. 4. Unrooted phylogenetic tree generated using the maximum likelihood optimality criterion and the preferred model of nucleotide evolution (SYM+ $\Gamma$ +I) selected by the hierarchical likelihood test and the AIC criterion.

Table 4  
Log likelihood scores ( $-\ln L$ ) of models calculated using the single NJ tree topology used in the hLRT

Model	$-\ln L$	$\delta\text{AIC}$	Topology
JC	3813.934	455.742	Fig. 5a
JC+I	3652.954	135.782	Fig. 5a
JC+ $\Gamma$	3659.403	148.680	Fig. 5a
JC+I+ $\Gamma$	3650.366	132.606	Fig. 5a
F81	3811.594	455.062	Fig. 5a
F81+I	3651.362	136.598	Fig. 5a
F81+ $\Gamma$	3657.217	148.308	Fig. 5a
F81+I+ $\Gamma$	3648.494	132.862	Fig. 5a
K80	3768.922	367.718	Fig. 5b
K80+I	3602.037	35.948	Fig. 4
K80+ $\Gamma$	3606.883	45.640	Fig. 5a
K80+I+ $\Gamma$	3598.376	30.626	Fig. 5b
HKY	3766.357	368.588	Fig. 5b
HKY+I	3601.262	40.398	Fig. 4
HKY+ $\Gamma$	3605.544	48.962	Fig. 5b
HKY+I+ $\Gamma$	3597.331	34.536	Fig. 4
SYM	3743.752	325.378	Fig. 4
SYM+I	3583.904	7.682	Fig. 4
SYM+ $\Gamma$	3586.496	12.866	Fig. 5b
SYM+I+ $\Gamma$	3579.063	Best	Fig. 4
GTR	3737.487	318.848	Fig. 5b
GTR+I	3582.327	10.528	Fig. 4
GTR+ $\Gamma$	3583.892	13.658	Fig. 5b
GTR+I+ $\Gamma$	3576.966	1.806	Fig. 4

Significance of likelihood comparison summarized in Table 2. Topology reconstructed under each of 24 models representing various levels of complexity. See the caption of Fig. 3 for model references.

share a most recent common ancestor (Fig. 5a). Eight of the nine models that reproduce this topology share a common feature: they do not have a substitution matrix specifying different rates for substitutions between different nucleotide pairs. Seven other models reconstructed a tree in which frog and mammals formed a tetrapod clade and fishes formed a monophyletic group. However, in this tree, the pufferfish and medaka, generally considered derived fishes, are found at the root of the teleost clade, displacing the more primitive salmon and zebrafish (Fig. 5b). In total, eight models preferred the ‘correct’ topology; however no clear pattern of which models reconstruct the ‘correct’ tree exists for these data (Table 4). For example, not all models that include a parameter for among-site rate variation result in the ‘correct’ tree, and some models that are more complex than the best-fit model found the ‘correct’ tree and some reconstructed another topology. The lack of

a clear pattern in progression of model parameters and tree structure illustrates that it is often impossible to tell a priori which models will find the same tree as the best-fit model, a fact that underscores the importance of finding the best-fit model.

#### 4.4. Discussion

It is difficult to fully assess why some models reconstruct a topology inconsistent with generally accepted taxon relationships in this example, and multiple factors of sequence evolution are often the cause. In this case, the level of diversity may contribute to misleading results. A very high level of diversity means that many potential substitutions may be unaccounted for using simple models that consistently underestimate the number of substitutions for distantly related species [60]. Multiple substitutions at given sites may provide conflicting evidence for various relationships, weakening support for a clade or overall branching pattern. This lack of consistent support renders trees with different topologies statistically indistinguishable. We tested the statistical difference among trees using the Shimodaira–Hasegawa test [91], and found no significant difference between all three topologies (Fig. 5a,  $P=0.305$ ; Fig. 5b,  $P=0.572$ ). Since some more complicated models also fail to reconstruct the widely accepted ‘true’ phylogeny it is likely that other factors play a role in misleading phylogenetic analysis. For these data, other factors such as different rates of evolution in part of the tree may also decrease the usefulness of models.

Use of simplistic models in evolutionary genetics can be misleading if the sequences do not evolve according to a molecular clock [92]. We used a simple LRT to test whether or not these sequences evolve according to a molecular clock [5] to see if this may be a misleading factor for simpler models. The LRT statistic was 137 ( $P<0.0001$ ;  $df=7$ ) indicating that the model enforcing a strict molecular clock was a much worse fit to these data. These results corroborate those of Takezaki et al. [93], who found variable substitution rates among lineages of proteasome components. Most statistical models of nucleotide evolution are ‘stationary,’ in that model parameters are constant across the entire tree; however, non-stationary models have been formulated that allow

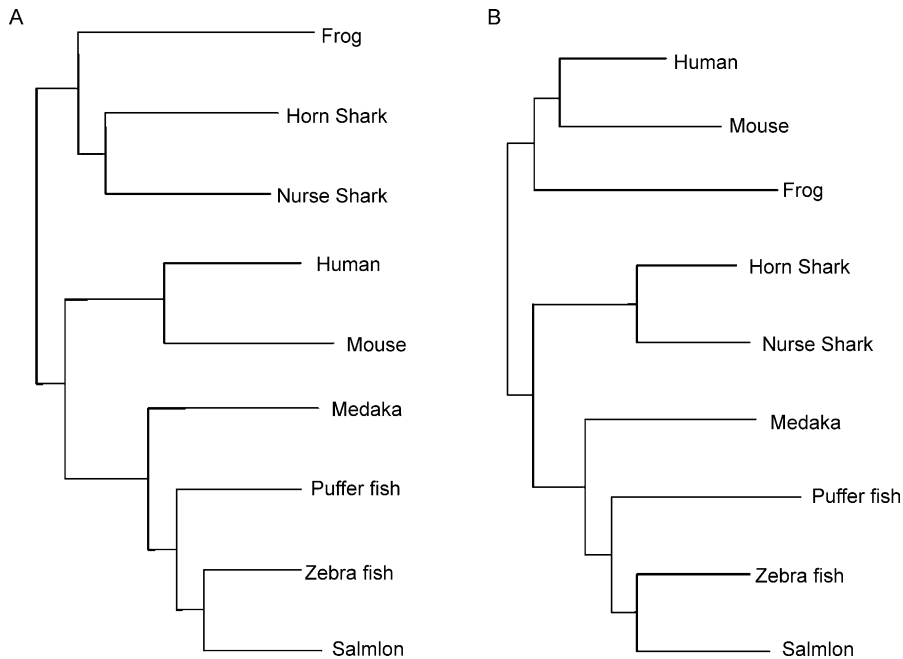


Fig. 5. Unrooted phylogenetic trees generated using the maximum likelihood optimality criterion. Twenty-four different models of nucleotide evolution (Table 4) were systematically selected to represent a range of models with differing levels of complexity, but arbitrarily selected with regard to how well they fit the data. These models were then applied to the data, and several of these models supported trees with topologies that differed from that reconstructed using the optimal model.

parameters to change with time [94,95]. Use of these models generally improves the fit to the data and performance of the method, but greatly increased model complexity. Here, the overall rate of evolution is different among branches of the tree, therefore this and other simplifying assumptions may affect model fitness and utility.

Other factors may be involved in the failure of some models, as Whelan et al. [96] indicate that positive or negative selection may be an unaccounted for dynamic that affects phylogenetic reconstructions. For instance, selective pressures can result in convergent evolution, causing divergent taxa to appear closely related. The test of model adequacy indicates support for the SYM +  $\Gamma$  + I model, but both models in that comparison make no provisions for natural selection and they both assume that data at each site is independent and identically distributed. Therefore, neither of these aspects of sequence evolution is evaluated in this comparison. Due to the coding nature of these sequences it is likely that both natural selection and non-independence of nucleotide

sites are prominent features of sequence evolution in these data.

The phylogenetics of proteasome components have been studied by others who included entire gene families in their sampling [93,97]. Previous phylogenetic work on proteasome components analyzed amino acid sequences, which can be an effective means of determining phylogeny in highly divergent data. These analyses employ a variety of methods including MP, a non-model based algorithm, and NJ with a Poisson corrected distance. (The Poisson amino acid model is analogous to the JC69 nucleotide model and assumes that all changes between amino acids occur at the same rate and all amino acids are found in equal frequency.) The JTT amino acid model [98] is also used to calculate maximum likelihood scores of three preset fixed topologies. The JTT model is more suited to the analysis of divergent amino acid sequences and is based on substitution rates in a large sample of related proteins [98]. In these cases, the use of a particular model is reported but no tests were conducted to select from a suite of available

amino acid models [99,100]. Statistical theory is often utilized through model-based phylogenetics, but the fuller potential and benefits of statistical analysis remains unemployed by not considering recent advancements in model selection. Many other examples of using evolutionary models for phylogenetic reconstruction without statistically evaluating the fit of a model are widespread in the literature [72].

Another example of differing trees obtained with differing phylogenetic methodologies can be found in a study of antigen receptors by Richards and Nelson [101]. They used two methods, MP and NJ, to reconstruct the evolutionary history of members of the immunoglobulin superfamily of genes using amino acid sequences. For NJ distance calculations, they do not specify which model of evolution was used to estimate genetic distances or mention how that model was selected. However, in their analysis the model-based NJ method outperformed the MP because more monophyletic clades reflected current immune receptor classifications established by function [101]. Even with a model of evolution, strong bootstrap support for many of the nodes in their analysis is lacking. Such a lack of support or conflicting trees may be expected when the natural limitations of protein size and ancient divergence constrain the size and signal of the sequence alignment used for analysis. Also, similar structures and function in families of genes can cause convergence at the molecular level. Finally, the period following the gene duplications that create multigene families is often marked with increased substitution rates or varying levels of natural selection [102]. The temporal and often temporary change in evolutionary process makes phylogenetic analysis with stationary models of evolution more difficult.

Other data sets will have different properties that play an important role in determining the best-fit model, and population data collected from a single species presents unique obstacles for evolutionary analysis. The phylogenetic methods discussed here are designed for use on hierarchically ordered data (each sampling unit has only a single ancestor) such as the creation of two species from one. Their use on population-based sampling from a single species presents other difficulties which may further complicate analysis and create misleading results, even with correct use of statistically justified models [103].

For instance, in a population sample, sequences may not be related in a hierarchical manner (each unit has two ancestors (parents) in a sexually reproducing species). Further, processes at the population level, such as recombination, result in the problem that different parts of a DNA sequence have different evolutionary histories, and cannot accurately be represented by a single phylogenetic tree [104]. Use of different parts of recombining trees typically leads to different trees that may not be correlated, depending on the relationship of the sequences that exchanged genetic information [105]. Recombination also alters estimates of mutation rates, dating of evolutionary events, and estimates of among-site variation [106,107]. Extra effort should be taken when using phylogenetic methods to analyze data from a single species to avoid pitfalls introduced by population-level processes, and methods designed for this purpose should be employed [103].

## 5. Summary

The estimation of phylogenetic trees or genetic distances is a complex statistical problem in which elements such as rate of evolution, branch length, and tree topology are represented by parameters in a model [56]. A phylogenetic tree and model parameters should be considered a hypothesis of evolutionary relationships statistically supported by particular data. It is important to ensure that any conclusions from evolutionary genetic analysis be as strongly supported as possible by using statistically relevant models. Results obtained using arbitrarily selected models may easily be contradicted simply by using different models that lend support to different hypotheses [52,55]. When model-based methods are used, their performance is optimized when the best model is used [37], thereby lending more credibility to results obtained using statistically justified models. Some estimate of the fit of the model to the data should be calculated and used to select among available models rather than relying heavily on the robustness of the reconstruction method [72]. It is our position that statistical accuracy should not be sacrificed for the sake of ease or computational speed. Advancements in the statistics of model selection have already benefited every scientific

discipline that uses model-based analysis. Evolutionary genetic analysis is also experiencing similar progress from these advancements. New models and improved implementation, along with selection of models under a statistically rigorous framework will continue to enhance understanding of evolutionary patterns and processes underlying the variation found in genes of the immune system.

## Acknowledgements

We would like to thank Louis Du Pasquier and Ashley Sparrow for helpful comments on an earlier version of this manuscript. Janae Bos and Matt Walters provided helpful editorial and digital imagery assistance. David Bos was supported by the Marsden Fund of New Zealand and a PhD scholarship from the University of Canterbury. David Posada is supported by the 'Ramón y Cajal' program of the Spanish government.

## References

- [1] Page RDM, Holmes EC. *Molecular evolution: a phylogenetic approach*. Cambridge: Blackwell Science; 1998.
- [2] Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. *Nature Rev Genet* 2003;4:275–84.
- [3] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–25.
- [4] Fitch WM. Toward defining the course of evolution: minimal change for a specific tree topology. *Syst Zool* 1970;20:406–16.
- [5] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;17:368–76.
- [6] Larget B, Simon D. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol* 1999;16:750–9.
- [7] Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572–4.
- [8] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 2001;294:2310–4.
- [9] Brower A, DeSalle R, Vogler AP. Gene trees, species trees, and systematics: a cladistic perspective. *Ann Rev Ecol Syst* 1996;27:423–50.
- [10] Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann Rev Ecol Syst* 1997;28:437–66.
- [11] Nei M. Phylogenetic analysis in molecular evolutionary genetics. *Ann Rev Genet* 1996;30:371–403.
- [12] Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol* 2001;50:525–39.
- [13] Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. *Molecular systematics*. Sunderland, MA: Sinauer; 1996.
- [14] Li W-H. *Molecular evolution*. Sunderland, MA: Sinauer; 1997.
- [15] Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 1988;335:167–70.
- [16] Yang Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 1996;11:367–72.
- [17] Zhang J, Gu X. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* 1998;149:1615–25.
- [18] Gu X, Zhang J. A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol* 1997;14:1106–13.
- [19] Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. *Evolving genes and proteins*. New York: Academic Press; 1965. p. 97–166.
- [20] Kimura M. Evolutionary rate at the molecular level. *Nature* 1968;217:624–6.
- [21] Bromham L, Penny D. The modern molecular clock. *Nature Rev Genet* 2003;4:216–24.
- [22] Ayala FJ. Molecular clock mirages. *BioEssays* 1999;21:71–5.
- [23] Gillespie JH. *The causes of molecular evolution*. New York: Oxford University Press; 1991.
- [24] Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA* 1997;94:7799–806.
- [25] Merritt TJS, Quattro JM. Evidence for a period of directional selection following gene duplication in a neutrally expressed locus of Triosephosphate Isomerase. *Genetics* 2001;159:689–97.
- [26] Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. Estimating divergence times from molecular data on population genetic and phylogenetic time scales. *Ann Rev Ecol Syst* 2002;33:707–40.
- [27] Sorhannus U, Van Bell C. Testing for equality of molecular evolutionary rates: a comparison between a relative-rate test and a likelihood ratio test. *Mol Biol Evol* 1999;16:849–55.
- [28] Graur D, Martin W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet* 2004;20:80–6.
- [29] Huelsenbeck JP, Larget B, Swofford DL. A compound process for relaxing the molecular clock. *Genetics* 2000;154:1879–92.
- [30] Sanderson MJ. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* 1997;14:1218–32.
- [31] Yoder AD, Yang Z. Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol* 2000;17:1081–90.

- [32] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 1994;11:459–68.
- [33] Huelsenbeck JP, Hillis DM. Success of phylogenetic methods in the four-taxon case. *Syst Biol* 1993;42:247–64.
- [34] Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 1978;27:401–10.
- [35] Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of the maximum likelihood over neighbor joining. *Mol Biol Evol* 1995;12:843–9.
- [36] Yang Z. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst Biol* 1994;43:329–42.
- [37] Huelsenbeck JP. Performance of phylogenetic methods in simulation. *Syst Biol* 1995;44:17–48.
- [38] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism*. New York, USA: Academic Press; 1969. p. 21–132.
- [39] Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16:111–20.
- [40] Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci USA* 1981;78:454–8.
- [41] Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lec Math Life Sci* 1986;17:57–86.
- [42] Hasegawa M, Kishino H, Yano T. Dating the human–ape split by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985;22:160–74.
- [43] Felsenstein J. *PHYLIP (Phylogenetic inference package)*. Seattle, WA: University of Washington; 1995.
- [44] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in Humans and Chimpanzees. *Mol Biol Evol* 1993;10:512–26.
- [45] Sullivan J, Holsinger KA, Simon C. Among site rate variation and phylogenetic analysis of 12s rRNA in Sigmontine rodents. *Mol Biol Evol* 1995;12:988–1001.
- [46] Yang Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* 1993;10:1396–401.
- [47] Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 1994;39:306–14.
- [48] Pedersen A-MK, Wiuf C, Christiansen FB. A codon-based model designed to describe lentiviral evolution. *Mol Biol Evol* 1998;15:1069–81.
- [49] Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 1994;11:715–24.
- [50] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 1994;11:725–36.
- [51] Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon substitution models for heterogeneous selection pressure and amino acid sites. *Genetics* 2000;155:431–49.
- [52] Kelsey CR, Crandall KA, Voevodin AF. Different models, different trees: the geographic origin of PTLV-I. *Mol Phylogent Evol* 1999;13:336–47.
- [53] Gu X, Li W-H. A general additive distance with time-reversibility and rate variation among nucleotide sites. *Proc Natl Acad Sci USA* 1996;93:4671–6.
- [54] Buckley TR, Simon C, Chambers GK. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst Biol* 2001;50:67–86.
- [55] Cunningham CW, Zhu H, Hillis DM. Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 1998;52:978–87.
- [56] Yang Z, Goldman N, Friday A. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst Biol* 1995;44:384–99.
- [57] Wakeley J. Substitution rate variation among sites and the estimation of transition bias. *Mol Biol Evol* 1994;11:436–42.
- [58] Tajima F, Takezaki N. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Mol Biol Evol* 1994;11:278–86.
- [59] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition–transversion and G+C content biases. *Mol Biol Evol* 1992;9:678–87.
- [60] Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol* 1994;11:316–24.
- [61] Sanderson MJ, Kim J. Parametric phylogenetics? *Syst Biol* 2000;49:817–29.
- [62] Sullivan J, Swofford DL. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst Biol* 2001;50:723–9.
- [63] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;39:783–91.
- [64] Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol* 1993;42:182–92.
- [65] Buckley TR, Cunningham CW. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol Biol Evol* 2002;19:394–405.
- [66] Lemmon AR, Moriarity EC. The importance of proper model assumption in Bayesian Phylogenetics. *Syst Biol* 2004;53:265–77.
- [67] Wilcox TP, Zwickl DJ, Heath TA, Hillis DM. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol Phylogent Evol* 2002;25:361–71.
- [68] Suzuki Y, Glazko GV, Nei M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci USA* 2002;99:16138–43.
- [69] Simmons MP, Pickett KM, Miya M. How meaningful are Bayesian support values? *Mol Biol Evol* 2004;21:188–99.

- [70] Takahashi K, Nei M. 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* 2000;17:1251–8.
- [71] Burnham KP, Anderson DR. Model selection and multi-model inference: a practical information-theoretic approach. New York: Springer; 2002.
- [72] Posada D, Crandall KA. Selecting the best-fit model of nucleotide substitution. *Syst Biol* 2001;50:580–601.
- [73] Swofford DL. PAUP\* phylogenetic analysis using parsimony (\*and other methods). Version 4.0. Sunderland, MA: Sinauer; 1998.
- [74] Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. Tempe: Arizona State University; 2001.
- [75] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001;17:754–5.
- [76] Rzhetsky A, Nei M. Tests of applicability of several substitution models for DNA sequence data. *Mol Biol Evol* 1995;12:131–51.
- [77] Whelan S, Goldman N. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol* 1999;16:1292–9.
- [78] Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics. *Syst Biol* 2000;49:652–70.
- [79] Goldman N, Whelan S. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol Biol Evol* 2000;17:975–8.
- [80] Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* 2001;98:13757–62.
- [81] Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Contr* 1974;19:716–23.
- [82] Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol* 2001;18:1001–13.
- [83] Huelsenbeck JP, Larget B, Alfaro ME. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol* 2004;21:1123–33.
- [84] Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–4.
- [85] Raftery AE. Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. *Markov chain Monte Carlo in practice*. London: Chapman & Hall; 1996. p. 163–87.
- [86] Navidi WC, Churchill GA, von Haeshler A. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol Biol Evol* 1991;8:128–43.
- [87] Goldman N. Statistical tests of models of DNA substitution. *J Mol Evol* 1993;36:182–98.
- [88] Bollback JP. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* 2002;19:1171–80.
- [89] Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 1998;14:817–8.
- [90] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–80.
- [91] Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 1999;16:1114–6.
- [92] Rzhetsky A, Sitnikova T. When is it safe to use an oversimplified substitution model in tree making? *Mol Biol Evol* 1996;13:1255–65.
- [93] Takezaki N, Zaleska-Rutczynska Z, Figueroa F. Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. *Gene* 2002;282:179–87.
- [94] Gu X, Li W-H. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc Natl Acad Sci USA* 1998;95:5899–905.
- [95] Huelsenbeck JP, Nielsen R. Variation in the pattern of nucleotide substitution across sites. *J Mol Evol* 1999;48:86–93.
- [96] Whelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 2001;17:262–72.
- [97] Hughes AL. Evolution of the proteasome components. *Immunogenetics* 1997;46:82–92.
- [98] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comp Appl Bioscience* 1992;8:275–82.
- [99] Kishino H, Miyata T, Hasegawa M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol* 1990;31:151–60.
- [100] Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 2001;18:691–9.
- [101] Richards MH, Nelson JL. The evolution of vertebrate antigen receptors: a phylogenetic approach. *Mol Biol Evol* 2000;17:146–55.
- [102] Moore RC, Purugganan MD. The early stages of duplicate gene evolution. *Proc Natl Acad Sci USA* 2003;100:15682–7.
- [103] Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol* 2001;16:37–45.
- [104] Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 2000;156:879–91.
- [105] Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* 2002;54:396–402.
- [106] Satta Y, Kupferman H, Li Y-J, Takahata N. Molecular clock and recombination in primate MHC genes. *Immunol Rev* 1999;167:367–79.
- [107] Schierup MH, Mikkelsen AM, Hein J. Recombination, balancing selection, and phylogenies in MHC and self-incompatibility genes. *Genetics* 2001;159:1833–44.
- [108] Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol* 1994;39:315–29.
- [109] Rodriguez F, Oliver JF, Marin A, Medina JR. The general stochastic model of nucleotide substitution. *J Theor Biol* 1990;142:485–501.