10

Selecting models of evolution

THEORY

David Posada

10.1 Models of evolution and phylogeny reconstruction

Phylogenetic reconstruction is a problem of statistical inference. Since statistical inferences cannot be drawn in the absence of probabilities, the use of a model of nucleotide substitution or amino acid replacement – a ***model of evolution*** – becomes indispensable when using DNA or protein sequences to estimate phylogenetic relationships among taxa. Models of evolution are sets of assumptions about the process of nucleotide or amino acid substitution (see Chapters 4 and 9). They describe the different probabilities of change from one nucleotide or amino acid to another along a phylogenetic tree, allowing us to choose among different phylogenetic hypotheses to explain the data at hand. Comprehensive reviews of models of evolution are offered elsewhere (Swofford *et al.*, 1996; Liò & Goldman, 1998).

As discussed in the previous chapters, phylogenetic methods are based on a number of assumptions about the evolutionary process. Such assumptions can be implicit, like in ***parsimony*** methods (see Chapter 8), or explicit, like in distance or ***maximum likelihood*** methods (see Chapters 5 and 6, respectively). The advantage of making a model explicit is that the parameters of the model can be estimated. Distance methods can only estimate the number of substitutions per site. However, maximum likelihood methods can estimate all the relevant parameters of the model of evolution. Parameters estimated via maximum likelihood have desirable statistical properties: as sample sizes get large, they converge to the true value and

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge University Press.

have the smallest possible variance among all estimates with the same expected value.

It is well known that the use of one model of evolution or another may change the results of a phylogenetic analysis (Sullivan & Joyce, 2005). When the model assumed is wrong, branch lengths, ***transition/transversion*** ratio, and divergence may be underestimated, while the strength of rate variation among sites may be overestimated. Simple models tend to suggest that a clade is significantly supported when it cannot be, and tests of evolutionary hypotheses (e.g. of the ***molecular clock***, see Chapter 11) can become conservative. In general, phylogenetic methods may be less accurate (recover an incorrect tree more often), or may be inconsistent (converge to an incorrect tree with increased amounts of data) when the assumed model of evolution is wrong. Cases where the use of wrong models increases phylogenetic performance are the exception, and they rather represent a bias towards the true tree due to violated assumptions. Indeed, models are not important just because of their consequences in the phylogenetic analysis, but because the characterization of the evolutionary process at the molecular level is itself relevant.

Models of evolution make assumptions to make complex problems computationally tractable. A model becomes a powerful tool when, despite its simplified assumptions, it can fit the data and make accurate predictions about the problem at hand. The performance of a method is maximized when its assumptions are satisfied, and some indication of the fit of the data to the phylogenetic model is necessary. If the model used may influence the results of the analysis, it becomes crucial to decide which is the most appropriate model to work with.

Before proceeding further, a word of caution should be said when selecting best-fit models for heterogeneous data, for example, when joining different genes for the phylogenetic analysis, or coding and non-coding regions. Since different genomic regions are subjected to different selective pressures and evolutionary constraints, a single model of evolution may not fit well with all the data. Nowadays, some options exist for a combined analysis in which each data partition (e.g. different genes) has its own model (Nylander *et al.*, 2004). In addition, *mixture models* consider the possibility of the model varying in different parts of the alignment (Pagel & Meade, 2004).

10.2 Model fit

In general, models that are more complex will fit the data better than simpler ones just because they have more parameters. An *a priori* attractive procedure to select a model of evolution would be the arbitrary use of the most complex, parameter-rich model available. However, when using complex models a large

number of parameters need to be estimated, and this has several disadvantages. First, the analysis becomes computationally difficult, and requires a large amount of time. Second, as more parameters need to be estimated from the same amount of data, more error is included in each estimate. Ideally, it would be advisable to incorporate as much complexity as needed, i.e. to choose a model that is intricate enough to explain the data, but not that complicated that requires impractical long computations or large data sets to obtain accurate estimates.

The best-fit model of evolution for a particular data set can be selected using sound statistical techniques. During the last few years different approaches have been proposed, to select the best-fit model of evolution within a collection of candidate models, like *hierarchical* ***likelihood ratio tests*** *(hLRTs), information criteria, Bayesian,* or *performance-based approaches.* In addition, although not considered here, the overall adequacy of a particular model can also be evaluated using different procedures (Goldman, 1993; Bollback, 2002).

Regardless of the model selection strategy chosen, the fit of a model can be measured through the ***likelihood function.*** The ***likelihood*** is proportional to the probability of the data (D) given a model of evolution (M), a vector of K model parameters (θ), a tree topology (τ), and a vector of branch lengths (ν):

$$L = P(D \mid M, \theta, \tau, \nu) \tag{10.1}$$

When the goal is to compute the likelihood of a model, the parameter values and the tree affect the calculations, but they are not really what we want to infer (they are ***nuisance parameters***). A standard strategy to "remove" nuisance parameters is to utilize their ***maximum likelihood estimates*** (***MLEs***), which are the values that make the likelihood function as large as possible:

$$\hat{\theta}, \hat{\tau}, \hat{\nu} = \max_{\theta, \tau, \nu} L(\theta, \tau, \nu) \tag{10.2}$$

Note that, to facilitate the computation, we usually work with the maximized log likelihood:

$$\ell = \ln P(D \mid M, \hat{\theta}, \hat{\tau}, \hat{\nu}) \tag{10.3}$$

Alternatively, in a Bayesian setting we can integrate the nuisance parameters out and obtain the *marginal probability* of the data given only the model ($P(D|M)$, also called *model likelihoods*), typically using computationally intensive techniques like ***Markov chain Monte Carlo*** (***MCMC***). Integrating out the tree, branch lengths, and model parameters to obtain $P(D|M)$ is represented by:

$$P(D \mid M) = \int\int\int P(D \mid M, \theta, \tau, \nu) P(\theta, \tau, \nu \mid M)\, d\theta\, d\tau\, d\nu \tag{10.4}$$

10.3 Hierarchical likelihood ratio tests (hLRTs)

A standard way of comparing the fit of two models is to contrast their log likelihoods using the likelihood ratio test (LRT) statistic:

$$LRT = 2(\ell_1 - \ell_0) \tag{10.5}$$

where ℓ_1 is the maximum log likelihood under the more parameter-rich, complex model (alternative hypothesis) and ℓ_0 is the maximum log likelihood under the less parameter-rich simple model (null hypothesis). The value of this statistic is always equal to, or greater than, zero, even if the simple model is closest to the true model, simply because the superfluous parameters in the complex model provide a better explanation of the stochastic variation in the data than the simpler model. When the models compared are nested (i.e. the null hypothesis is a special case of the alternative hypothesis) and the null hypothesis is correct, this statistic is asymptotically distributed as a χ^2 distribution with a number of degrees of freedom equal to the difference in number of free parameters between the two models. When the value of the LRT is significantly large, the conclusion is that the inclusion of additional parameters in the alternative model increases the likelihood of the data significantly, and consequently the use of the more complex model is favored. On the other hand, a small difference in the log likelihoods indicates that the alternative hypothesis does not explain the data significantly better than the null hypothesis.

That two models are nested means that one model (null model or constrained model) is equivalent to a restriction of the possible values that one or more parameters can take in the other model (alternative, unconstrained or full model). For example, the Jukes–Cantor model (JC) (1969) and the Felsenstein (F81) (1981) models are nested. This is because the JC model is a special case of the F81, where the base frequencies are set to be equal (0.25), while in the F81 model these frequencies can be different. When comparing two different nested models through an LRT, we are testing hypotheses about the data. The hypotheses tested are those represented by the difference in the assumptions among the models compared. Several hypotheses can be tested hierarchically to select the best-fit model for the data set at hand among a set of possible models. Are the base frequencies equal? Is there a transition/transversion bias? Are all transition rates equal? Are there invariable sites? Is there rate homogeneity among sites? And so on. For example, testing the equal base frequencies hypothesis can be done with a LRT comparing JC vs. F81, as these models only differ in the fact that F81 allows for unequal base frequencies (alternative hypothesis), while JC assumes equal base frequencies (null hypothesis). Indeed, the same hypothesis could also have been evaluated by comparing JC + Γ vs. F81 + Γ, or K80 + I vs. HKY + I, or SYM vs. GTR. An example of such a

hierarchical LRT procedure (hLRT) for 24 models is shown in Fig. 10.1. The hLRTs can be easily accomplished by using the program MODELTEST (Posada & Crandall, 1998) for a set of 56 candidate models (see the practice section in this chapter).

10.3.1 Potential problems with the hLRTs

We should be aware that there are some potential problems derived from the use of pairwise LRTs for model selection (Posada & Buckley, 2004). The χ^2 distribution approximation for the *LRT* statistic may not be appropriate when the null model is equivalent to fixing some parameter at the boundary of its possible values (Whelan & Goldman, 1999). An example of this situation is the invariable sites test. In this case, the alternative hypothesis postulates that the proportion of invariable sites could range from 0 to 1. The null hypothesis (no invariable sites) is a special case of the alternative hypothesis, with the proportion of invariable sites fixed to 0, which is at the boundary of the range of the parameter in the alternative model. In this case, the use of a mixed χ^2 distribution (50% χ^2_0 and 50% χ^2_1) is appropriate. However, even after using the most appropriate χ^2 distribution, obtaining *correct* *P*-values for the LRT statistics can be difficult, because LRTs implicitly assume that at least one of the models compared is correct. Moreover, when the two competing hypotheses are not nested the χ^2 approximation may perform poorly when the data include very short sequences relative to the number of parameters to be estimated. In these cases, the null distribution of the LRT statistic can be approximated by ***Monte Carlo simulation***.

In addition, which model comparison is used to compare which hypothesis depends on the starting model of the hierarchy, and on the order in which different hypotheses are performed. For example, it could be possible to start with the simple JC or with the most-complex GTR + I + Γ. In the same way, a test for equal base frequencies could be performed first followed by a test for rate heterogeneity among sites, or vice versa. Many hierarchies of LRTs are possible, and they can result in different models being selected (Posada & Crandall, 2001; Posada & Buckley, 2004), and in some cases they can even lead to the estimation of different trees (Pol, 2004).

10.4 Information criteria

A different approach for model selection is the simultaneous comparison of all competing models. The idea again is to include as much complexity in the model as needed. To do that, the likelihood of each model is penalized by a function of the number of free parameters in the model (*K*); the more parameters, the bigger the penalty. The ***Akaike Information Criterion*** *or* ***AIC*** (Akaike, 1974) is an asymptotically unbiased estimator of the Kullback–Leibler information quantity

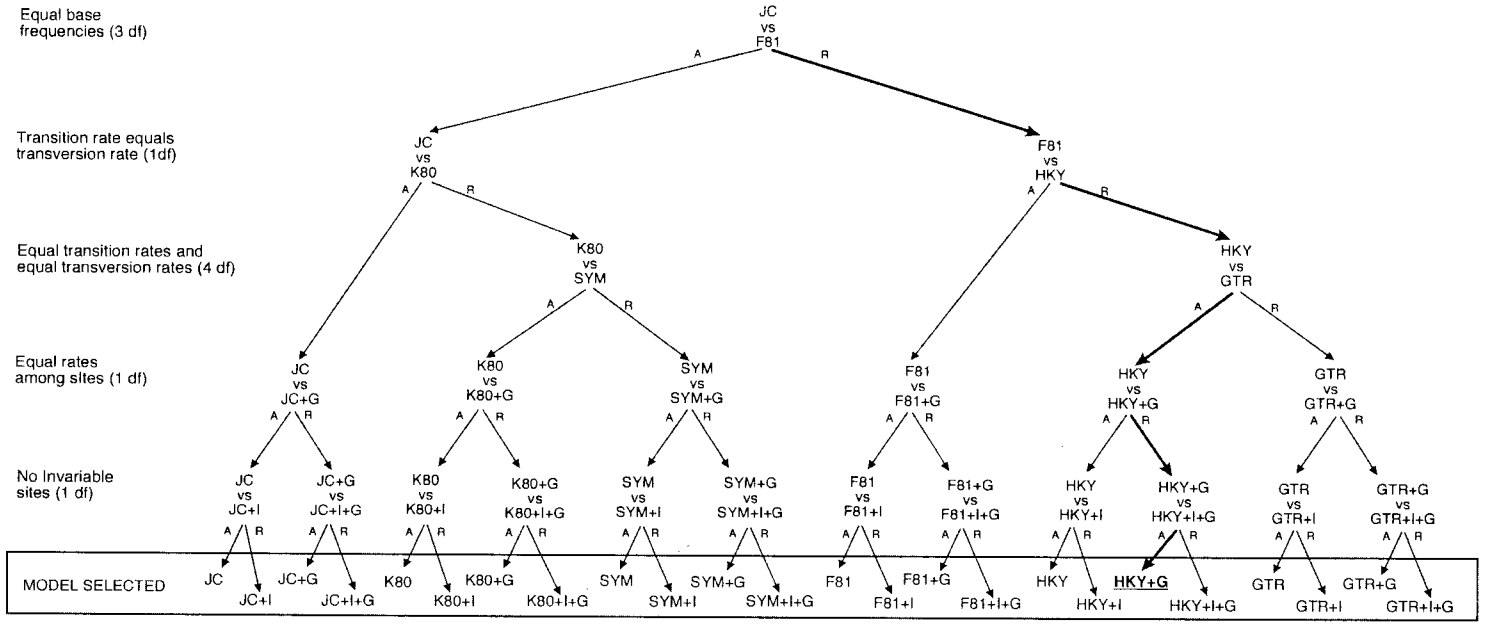

Fig. 10.1 Hierarchical likelihood ratio tests for 24 models of nucleotide substitution (in MODELTEST the number of models is 56; see Table 10.1). At each level the null hypothesis (model on top) is either accepted (A) or rejected (R). G: shape parameter of the gamma distribution; I: proportion of invariable sites. Here the model selected would be HKY+G as indicated by the pathway represented using bold arrows.

(Kullback & Leibler, 1951), which measures the expected distance between the true model and the estimated model:

$$AIC = -2\ell + 2K \tag{10.6}$$

We can think of the AIC as the amount of information lost when we use, say HKY85, to approximate the real process of molecular evolution. Hence, the model with the smallest AIC is preferred. If branch lengths are estimated *de novo* for every model, as is usually the case, K will include the number of branches (twice the number of taxa minus three). An advantage of the AIC is that it can be used to compare both nested and non-nested models. When sample size (n) is small compared with the number of parameters ($n/K < 40$) a corrected version of the AIC is recommended:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \tag{10.7}$$

Note that sample size is usually approximated by the total number of characters in the alignment, although what is the sample size of an alignment is still an open question. The AIC and AIC_c calculations are implemented in the programs ModelTest and ProtTest (Abascal *et al.*, 2005) for DNA and protein sequences, respectively.

10.5 Bayesian approaches

Model selection can be implemented in a Bayesian setting using ***Bayes factors***, ***posterior probabilities*** or the ***Bayesian Information Criterion***. Bayes factors are similar to the LRTs in that they compare the evidence (here, the model likelihoods) for two competing models:

$$B_{ij} = \frac{P(D|M_i)}{P(D|M_j)} \tag{10.8}$$

In this case, evidence for M_i is considered very strong if $B_{ij} > 150$, strong if $12 < B_{ij} < 150$, positive if $3 < B_{ij} < 12$, barely worth mentioning if $1 < B_{ij} < 3$, and negative (supports M_j) if $B_{ij} < 1$. Bayes factors for models of molecular evolution can be calculated using ***reversible jump MCMC*** (Huelsenbeck *et al.*, 2004) (see also Chapter 7).

In addition, when multiple models are considered, it is possible to choose the model with the highest ***posterior probability*** (Raftery, 1996). For R candidate models, the posterior probability of the ith model is:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{r=1}^{R} P(D|M_r)P(M_r)} \tag{10.9}$$

where $P(M)$ are the model ***prior probabilities.***

Both Bayes factors and model posterior probabilities can be difficult to compute. The Bayesian Information Criterion (BIC) (Schwarz, 1978) provides an approximate solution to the natural log of the Bayes factor:

$$BIC = -2\ell + K \log n \tag{10.10}$$

The smaller the BIC, the better the fit of the model to the data. Given equal priors for all competing models, choosing the model with the smallest BIC is equivalent to selecting the model with the maximum posterior probability. Because with standard alignments the natural log of n is usually >2, the BIC tends to choose simpler models than the AIC. Like the AIC, the BIC can be used to compare nested and non-nested models. The BIC calculation is implemented in the programs MODELTEST and PROTTEST.

10.6 Performance-based selection

Arguing that there is no guarantee that the best-fit models will produce the best estimates of phylogeny, Minin *et al.* (2003) developed a novel approach that selects models on the basis of their phylogenetic performance, measured as the expected error on branch lengths estimates weighted by their BIC. Under this *decision theoretic* framework (DT) the best model is the one with that minimizes the risk function:

$$C_i \approx \sum_{j=1}^{R} \left\| \hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j \right\| \frac{e^{-BIC_i/2}}{\sum_{j=1}^{R} e^{-BIC_i/2}} \tag{10.11}$$

where

$$\left\| \hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j \right\|^2 = \sum_{l=1}^{2t-3} (\hat{B}_{il} - \hat{B}_{jl})^2 \tag{10.12}$$

and where t is the number of taxa.

Indeed, simulations suggested that models selected with this criterion result in slightly more accurate branch length estimates than those obtained under models selected by the hLRTs (Minin *et al.*, 2003; Abdo *et al.*, 2005).

10.7 Model selection uncertainty

One big advantage of the AIC, Bayesian, and DT methods over the hLRTs is that they can rank models, allowing us to assess how confident we are in the model selected. Indeed, models could be ranked according to posterior probabilities, and credible intervals could easily be constructed by summing these probabilities. For

other relative measures like the AIC or BIC, we could present their *differences* (Δ). For example, for the *i*th model, the AIC (or BIC) difference is:

$$\Delta AIC_i = AIC_i - \min AIC \tag{10.13}$$

where min *AIC* is the smallest ***AIC*** value among all candidate models.

Very conveniently, we can use these *differences* to obtain the relative weight (w_i) of each model:

$$w_i = \frac{\exp(-1/2\Delta_i)}{\sum_{r=1}^{R} \exp(-1/2\Delta_r)} \tag{10.14}$$

Note that the weights for every model add to 1, so it is easy to establish a 95% confidence set of models for the best models by summing the weights from largest to smallest from largest to smallest until the sum is just 0.95 (or similar).

10.8 Model averaging

Very interestingly, the model weights (or the posterior probabilities) allow us to obtain a *model-averaged* estimate (also called a *multimodel* estimate) of any parameter (Raftery, 1996; Wasserman, 2000; Burnham & Anderson, 2003; Hoeting *et al.*, 1999; Madigan & Raftery, 1994). For example, a model-averaged estimate of the relative substitution rate between adenine and cytosine (φ_{A-C}) using the model weights (w) for *R* candidate models would be:

$$\hat{\bar{\varphi}}_{A-C} = \frac{\sum_{i=1}^{R} w_i I_{\varphi_{A-C}}(M_i)\varphi_{A-C_i}}{w + (\varphi_{A-C})}, \tag{10.15}$$

where

$$w + (\varphi_{A-C}) = \sum_{i=1}^{R} w_i I_{\varphi_{A-C}}(M_i), \tag{10.16}$$

and

$$I_{\varphi_{A-C}}(M_i) = \begin{cases} 1 & \text{if } \varphi_{A-C} \text{ is in model } M_i, \\ 0 & \text{otherwise} \end{cases} \tag{10.17}$$

Remarkably, it is possible to construct a model-averaged estimate of the phylogeny itself. Also, note that some parameters do not have the same interpretation across different models. For example, the shape of the gamma distribution (α, commonly used to describe among-site rate variation) in a + Γ model is not the same parameter as α in a + I + Γ model.

Furthermore, if we sum up the weights of the models that contain a given parameter we will get an estimate of the *relative importance* of that parameter (ranging

from 0 to 1). For example, the relative importance of the relative substitution rate between adenine and cytosine is simply the $w + (\varphi_{A-C})$ coefficient above. Because we usually do not explore all the possible combinations of parameters in the set of candidate models, the relative importance of some parameters can be correlated.