

# Selecting models of evolution

## THEORY

David Posada

### 10.1 Models of evolution and phylogeny reconstruction

---

Phylogenetic reconstruction is regarded as a problem of statistical inference. Because statistical inferences cannot be drawn in the absence of a probability model, the use of a model of nucleotide or amino-acid substitution – an *evolutionary model* – becomes necessary when using DNA or amino-acid sequences to estimate phylogenetic relationships among organisms. Evolutionary models are sets of assumptions about the process of nucleotide or amino-acid substitution (see Chapters 4 and 8). They describe the different probabilities of change from one nucleotide or amino acid to another, with the aim of correcting for unseen changes along the phylogeny. Although this chapter focuses on *models of nucleotide substitution*, all the points made herein can be applied directly to models of amino-acid replacement. Comprehensive reviews of models of evolution are offered by Swofford et al. (1996) and Lió and Goldman (1998).

As discussed in the previous chapters, the methods used in molecular phylogeny are based on a number of assumptions about how the evolutionary process works. These assumptions can be implicit, like in *parsimony methods* (see Chapter 7), or explicit, like in *distance* or *maximum-likelihood methods* (see Chapters 5 and 6). The advantage of making a model explicit is that the parameters of the model may be estimated. Distance methods may estimate only from the data of a single parameter of the model – the number of substitutions per site. However, maximum likelihood can estimate all the relevant parameters of the *substitution model*. Parameters estimated via maximum likelihood have desirable statistical properties: as sample sizes get large, they converge to the true parameter value and have the smallest possible variance among all estimates with the same expected value. Most important, as shown in the following sections, maximum likelihood provides a framework in

which different evolutionary hypotheses can be statistically tested rigorously and objectively.

## 10.2 The relevance of models of evolution

---

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

Evolutionary models are always simplified, and they often make assumptions just to turn a complex problem into a computationally tractable one. 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. Unfortunately – and despite their relevance – the unjustified use of evolutionary models is still a common practice in phylogenetic studies. If the model used may influence results of the analysis, it becomes crucial to decide which is the most appropriate model with which to work.

## 10.3 Selecting models of evolution

---

In general, more complex models fit the data better than simpler ones. An a priori attractive procedure to select a model of evolution is the arbitrary use of complex, parameter-rich models. However, when using complex models, numerous parameters need to be estimated, which has several disadvantages. First, the analysis becomes computationally difficult, and requires significant 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; that is, to choose a model complex enough to explain the

data but not so complex that it 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 through statistical testing. The fit to the data of different models can be contrasted through *likelihood ratio tests (LRTs)* or *information criteria* to select the best-fit model within a set of possible ones. In addition, the overall adequacy of a particular model to fit the data can be tested using an LRT.

A word of caution is necessary when selecting best-fit models for heterogeneous data; for example, when joining different genes for the phylogenetic analysis or a coding and a noncoding region. Because different genomic regions are subjected to different selective pressures and evolutionary constraints, a single substitution model may not fit well all the data. Although some options exist for the combined analysis of multiple-sequence data (Yang, 1996; Salemi, Desmyter, and Vandamme, 2000a), these are computationally expensive. An alternative solution would be to run separate analyses for each gene or region.

#### 10.4 The likelihood ratio test

In Chapter 6, the *likelihood function* was introduced as the conditional probability of the data (i.e., aligned *homologous* sequences) given the following hypothesis (i.e., a model of substitution with a set of parameters  $\theta$  – for example, base frequencies or transition/transversion ratio – and the tree  $\tau$ , including branch lengths):

$$\begin{aligned} L(\tau, \theta) &= \text{Prob}(\text{Data} \mid \tau, \theta) \\ &= \text{Prob}(\text{Aligned sequences} \mid \text{tree, model of evolution}) \end{aligned} \quad (10.1)$$

with *maximum-likelihood estimates (MLEs)* of  $\tau$  and  $\theta$  making the likelihood function as large as possible:

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

A natural way of comparing two models is to contrast their likelihoods using the LRT statistic:

$$\Delta = 2(\log_e L_1 - \log_e L_0) \quad (10.3)$$

where  $L_1$  is the maximum likelihood under the more parameter-rich, complex model (i.e., alternative hypothesis) and  $L_0$  is the maximum likelihood under the less parameter-rich, simple model (i.e., null hypothesis). The value of this statistic is always equal to or greater than zero – even if the simple model is the true one – 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  $\chi^2$ , with a number of degrees of freedom equal to the difference in number of free parameters between the two models. In other words, the number of degrees of freedom is the number of restrictions on the parameters of the alternative hypothesis required to derive the particular case of the null hypothesis. When the value of the LRT is significant (i.e.,  $<0.05$  or  $<0.01$ ), the conclusion is that the inclusion of additional parameters in the alternative model significantly increases the likelihood of the data and, consequently, the use of the more complex model is favored. Conversely, a difference in the log likelihood close to zero means that the alternative hypothesis does not fit the data significantly better than the null hypothesis (i.e., adding those particular parameters to the null model does not give a better explanation of the data).

That two models are nested means that one model (i.e., null model or constrained model) is equivalent to restrict the possible values that one or more parameters can take in the other model (i.e., alternative, unconstrained, or full model). For example, the *Jukes-Cantor (JC) model* (1969) and the *Felsenstein (F81) model* (1981) are nested. This is because the JC model is a special case of the F81, where the base frequencies are set to be equal (all are 0.25); whereas in the F81 model, these frequencies can be different (e.g., 0.20., 0.60., 0.15, and 0.05). The  $\chi^2$  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 parameter space in the alternative model (Whelan and Goldman, 1999). An example of this situation is the invariable sites test, in which the alternative hypothesis postulates that the proportion of invariable sites could range from 0 to 1. The null hypothesis (i.e., 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  $\chi^2$  distribution (i.e., 50%  $\chi_0^2$  and 50%  $\chi_1^2$ ) is appropriate. Although the difference in likelihoods when comparing current models may be significant and the inaccuracy of the  $\chi^2$  approximation may not change results of these tests (Posada, 2001a; Posada and Crandall, 2001a), as more complex and realistic models are developed (in which the differences of likelihoods might be insignificant), the use of a mixed  $\chi^2$  distribution may be essential. The use of LRTs for hypothesis testing in phylogeny is reviewed by Huelsenbeck and Crandall (1997) and Huelsenbeck and Rannala (1997).

#### 10.4.1 LRTs and parametric bootstrapping

The  $\chi^2$  approximation to assess the significance of the LRT is not appropriate when the two competing hypotheses are not nested, and it may perform poorly when

the data include very short sequences relative to the number of parameters to be estimated. In this case, the null distribution of the LRT statistic can be approximated by the *Monte Carlo simulation*. The general strategy is as follows:

1. Select the competing models: one for the null hypothesis  $H_0$  and one for the alternative hypothesis  $H_1$ .
2. Estimate the tree and the parameters of the model under the null hypothesis.
3. Use the tree and the estimated parameters to simulate 200–1000 replicate data sets of the same size as the original.
4. For each simulated data set, estimate a tree and calculate its likelihood under the models representing  $H_0$  and  $H_1$  ( $L_0$  and  $L_1$ , respectively). Calculate the LRT statistic  $\Delta = 2 (\log_e L_1 - \log_e L_0)$ . These simulated  $\Delta$ s form the distribution of the LRT statistic if the null hypothesis was true (i.e., they constitute the null distribution of the LRT statistic).
5. The probability of observing the LRT statistic from the original data set if the null hypothesis is true is the number of simulated  $\Delta$ s bigger than the original  $\Delta$ , divided by the total number of simulated data sets. If this probability is smaller than a predefined value (usually 0.05),  $H_0$  is rejected.

The main disadvantage of parametric bootstrapping is its computational expensive-ness. Because the likelihood calculations must be repeated on each simulated data set, this approach becomes unfeasible when many sequences are considered, even for fast supercomputers. A general discussion on model-fitting through parametric bootstrapping can be found in Goldman (1993a and b). Huelsenbeck et al. (1996) provide an interesting review of the applications of parametric bootstrapping in molecular phylogenetics.

#### 10.4.2 Hierarchical LRTs

Comparing two different nested models through an LRT means 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. It is to our advantage to test one hypothesis at a time: Are the base frequencies equal? Is there a transition/transversion (ti/tv) bias? Are all transition rates equal? Are there invariable sites? Is there rate homogeneity among sites? For example, testing the equal-base-frequencies hypothesis can be done with a LRT comparing JC versus F81, because these models only differ in the fact that F81 allows for unequal base frequencies (i.e., alternative hypothesis), whereas JC assumes equal base frequencies (i.e., null hypothesis). However, the hypothesis also could be evaluated by comparing JC +  $\Gamma$  versus F81 +  $\Gamma$ , or K80 + I versus HKY + I, and so forth (see Chapter 4 for more details about the models). 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 +  $\Gamma$ . 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 some seem to be more effective in selecting the best-fit model (Posada, 2001a; Posada and Crandall, 2001a). An alternative to the use of a particular hierarchy of LRTs is the use of dynamical LRTs described in the next section. The main steps to perform the hierarchical LRTs are as follows:

1. Estimate a tree from the data (i.e., the *base tree*). This tree has been shown to not have influence in the final model selected as far as it is not a random tree (Posada and Crandall, 2001a). A *neighbor-joining* (NJ) tree will be fast and will do fine.
2. Estimate the likelihoods of the candidate models for the given data set and the base tree.
3. Compare the likelihoods of the candidate models through a hierarchy of LRTs (Figure 10.1) to select the best-fit model among the candidates.

The hierarchy of tests can be accomplished easily by using the program MODEL-TEST (Posada and Crandall, 1998).

### 10.4.3 Dynamical LRTs

An alternative to the use of a predefined hierarchy LRT is to let the data itself determine the order in which the hypotheses are tested. In this case, the hierarchy used does not have to be the same for different data sets. The algorithms suggested proceed as follows:

Algorithm 1 (bottom-up)

1. Start with the simplest model and calculate its likelihood. This is the current model.
2. Calculate the likelihood of the alternative models differing by one assumption and perform the corresponding nested LRTs.
3. If any hypotheses are rejected, the alternative model corresponding to the LRT with smallest associated *P-value* becomes the current model. In the case of several equally smallest p-values, select the alternative model with the best likelihood.
4. Repeat Steps 2 and 3 until the algorithm converges.

Algorithm 2 (top-down)

1. Start with the most complex model and calculate its likelihood. This is the current model.
2. Calculate the likelihood of the null models differing by one assumption and perform the corresponding nested LRTs.



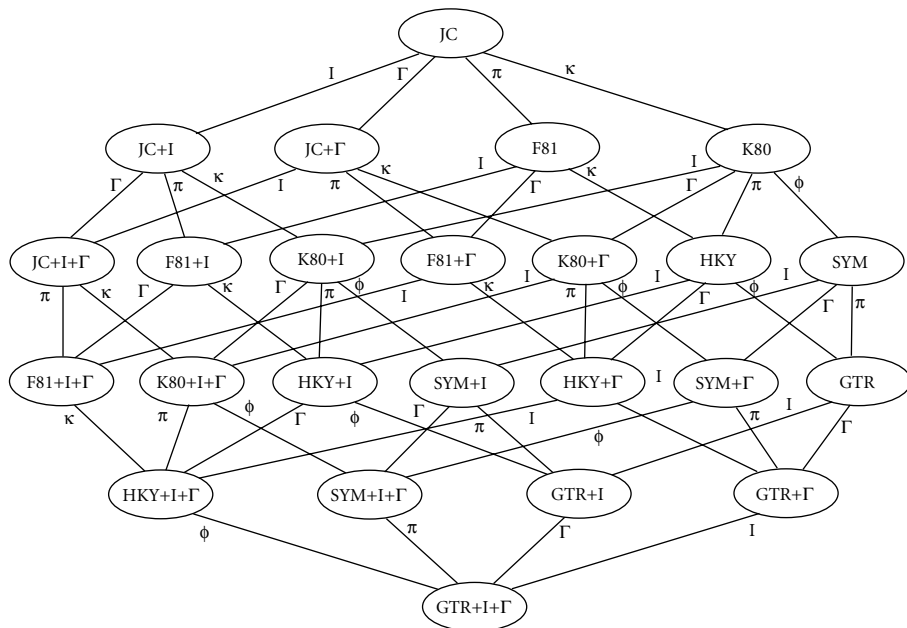


Figure 10.2 Dynamic LRTs. Starting with the simplest (JC) or the most complex model (GTR+I+ $\Gamma$ ), LRTs are performed among the current model and the alternative models that maximize the difference in likelihood.  $\pi$ : base frequencies;  $\kappa$ : transition/transversion bias;  $\phi$ : substitution rates among nucleotides;  $\Gamma$ : rate heterogeneity among sites;  $I$ : proportion of invariable sites.

3. If any hypotheses are not rejected, the null model corresponding to the LRT with the biggest associated p-value becomes the current model. In the case of several equally biggest p-values, select the null model with the best likelihood.
4. Repeat Steps 2 and 3 until the algorithm converges.

The alternative paths that the algorithm can generate can be represented graphically (Figure 10.2).

## 10.5 Information criteria

Whereas the LRTs compare two models at a time, 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 parameters in the model: the more parameters, the bigger the penalty. Two common information criteria are the *Akaike information criterion* (AIC) (Akaike, 1974) and the *Bayesian information criterion* (BIC) (Schwarz, 1974).

**10.5.1 AIC**

The AIC is an asymptotically unbiased estimator of the Kullback-Leibler information quantity (Kullback and Leibler, 1951), which measures the expected distance between the true model and the estimated model. The AIC takes into account not only the goodness of fit, but also the variance of the parameter estimates: the smaller the AIC, the better the fit of the model to the data. An advantage of the AIC is that it also can be used to compare both nested and non-nested models. It is computed as follows:

$$\text{AIC}_i = -2 \log_e L_i + 2 N_i \quad (10.4)$$

where  $N_i$  is the number of free parameters in the  $i$ th model and  $L_i$  is the maximum-likelihood value of the data under the  $i$ th model. The AIC calculation is implemented in the program MODELTEST.

**10.5.2 BIC**

The BIC provides an approximate solution to the natural log of the Bayes factor, especially when sample sizes are large and competing hypotheses are nested (Kass and Wasserman, 1994). The Bayes factor measures the relative support that data gives to different models; however, its computation often involves difficult integrals and an approximation becomes convenient. Like the AIC, the BIC can be used to compare nested and nonnested models. Its definition is as follows:

$$\text{BIC}_i = -2 \log_e L_i + N_i \log_e n \quad (10.5)$$

where  $n$  is the sample size (sequence length): the smaller the BIC, the better the fit of the model to the data. Because in real data analysis the natural log of  $n$  is usually greater than 2, the BIC should tend to choose simpler models than the AIC.

**10.6 Fit of a single model to the data**

Once a model has been shown to offer a better fit than other models, it is important to assess its general adequacy to the data. To do that, an upper bound to which the likelihood of any model can be compared is needed. This upper bound corresponds to an unconstrained model of evolution, and can be estimated by viewing the sites (i.e., columns) of an alignment as a multinomial sample. The likelihood function under the multinomial distribution for  $n$  aligned DNA sequences of length  $N$  sites (excluding gapped sites) has the form:

$$L = \prod_{b \in \mathfrak{B}} (p_b)^{n_b} \quad (10.6)$$

where  $\mathfrak{N}$  is a set of  $4^n$  possible nucleotide patterns that may be observed at each site,  $p_b$  is the probability that any site exhibits the pattern  $b$  in  $\mathfrak{N}$  given the tree and a substitution model, and  $n_b$  is the number of times the pattern  $b$  is observed out of the  $N$  sites. This comparison provides an idea of how well a particular model explains the observed data. However, this test is very stringent, and most models are usually rejected against the multinomial model. This does not mean that current models are inadequate to provide reasonable estimates, but rather that current models do not provide a perfect description of the underlying evolutionary process. Because a model of evolution is never expected to be correct in every detail, this test is perhaps best used to estimate how far the assumed model deviates from the underlying process that generated the data (Swofford et al., 1996).

Rzhetsky and Nei (1995) also developed several tests using linear invariants for the applicability of a particular model to the data. They tested whether the deviation from the expected invariant would be significant if the evaluated model were true. Although these tests do not require the use of an initial phylogeny and they are independent of evolutionary time, they are model-specific; currently, they can be applied only to a small set of substitution models.

## 10.7 Testing the molecular clock hypothesis

Between 1962 and 1965, before Kimura postulated the neutral theory of evolution (Kimura, 1968), Zuckerkandl and Pauling published two fundamental papers on the *evolutionary rate* of proteins (Zuckerkandl and Pauling 1962 and 1965). They noticed that the *genetic distance* of two sequences coding for the same protein, but isolated from different species, seems to increase linearly with the divergence time of the two species. Because several proteins showed a similar behavior, Zuckerkandl and Pauling hypothesized that the rate of evolution for any given protein is constant over time. This suggestion implies the existence of a type of molecular clock ticking faster or slower for different genes, but at a more or less constant rate for any given gene among different phylogenetic lineages (Figure 10.3). The hypothesis received almost immediate popularity for several reasons. If a molecular clock exists and the rate of evolution of a gene can be calculated, then this information can easily be used for dating the unknown divergence time between two species just by comparing their DNA or protein sequences. Conversely, if the information about the divergence time between two species (e.g., estimated from fossil data) is known, then the rate of molecular evolution of a given gene can be inferred. Moreover, phylogeny reconstruction is much easier and more accurate under the assumption of a molecular clock (see Chapter 5).

The molecular-clock hypothesis is in perfect agreement with the neutral theory of evolution (Kimura, 1968 and 1983). In fact, the existence of a clock seems to be

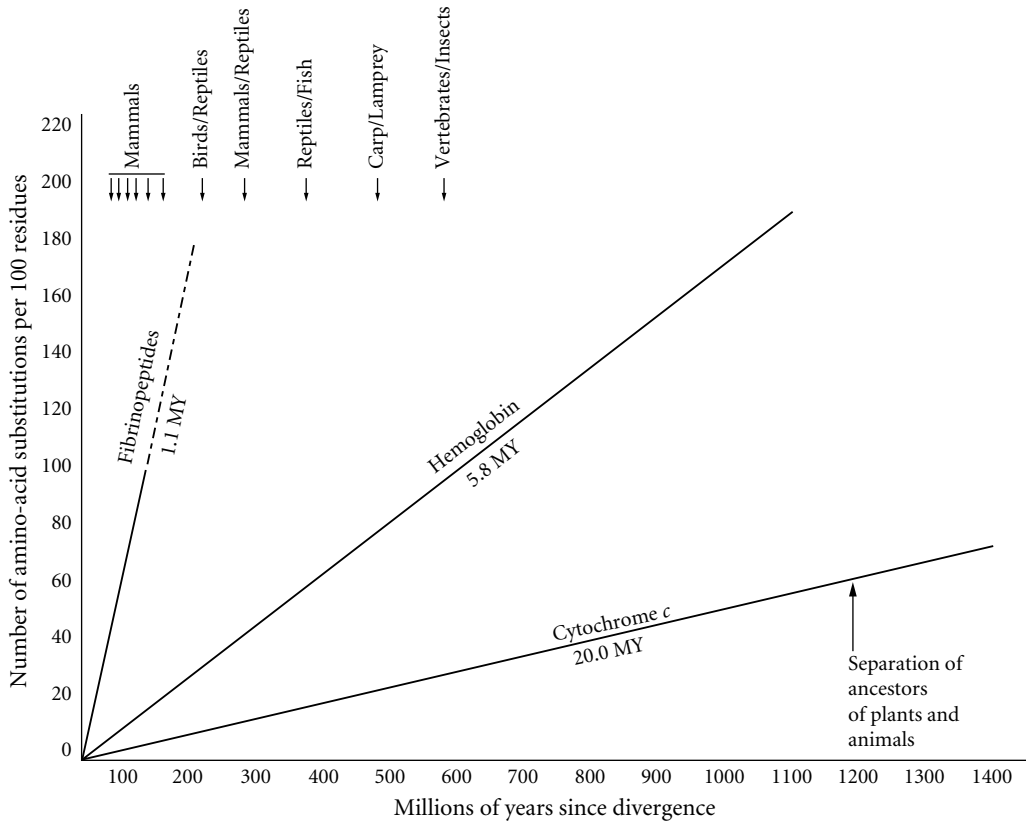


Figure 10.3 Molecular clock ticking at different speed in different proteins. Fibrinopeptides are relatively unconstrained and have a high neutral substitution rate, whereas cytochrome *c* is more constrained and has a lower neutral substitution rate (after Hartl and Clark, 1997).

a major support of the neutral theory against natural selection (see Chapter 1). A detailed discussion of the molecular clock is beyond the scope of this book. Excellent reviews can be found in textbooks of molecular evolution (e.g., Hillis et al., 1996; Li, 1997; Page and Holmes, 1998). The next section focuses more on how to test the clock hypothesis for a group of *taxa* with known phylogenetic relationships.

### 10.7.1 The relative rate test

According to the molecular-clock hypothesis, two taxa that shared a common ancestor  $t$  years ago should have accumulated more or less the same number of substitutions during time  $t$ . In most cases, however, the ancestor is unknown and there is no possibility to directly test the constancy of the evolutionary rate. The problem can be solved by considering an **outgroup**: that is, a more distantly related species (Figure 10.4). Under a perfect molecular clock,  $d_{AO}$  – the number of substitutions between taxon A and the outgroup – is expected to be equal to  $d_{BO}$  – the number of

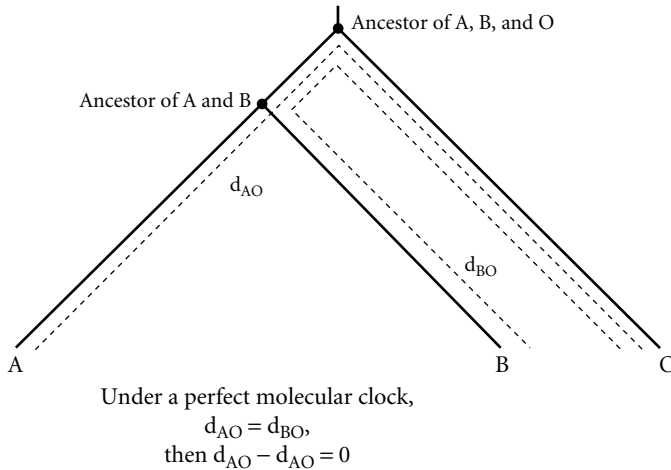


Figure 10.4 The relative rate test. Under a molecular clock, the distance from A to O should be the same as the distance from B to O.

substitutions between taxon B and the outgroup. The relative rate test evaluates the molecular clock hypothesis comparing whether  $d_{AO} - d_{BO}$  is significantly different from zero. When this is the case, the sign of the difference indicates which taxon is evolving faster or slower. The relative rate test assumes that the phylogenetic relationships among the taxa are known, which makes the test problematic for taxa, such as the placental mammals with still uncertain phylogeny. In these cases, it would not be a good idea to choose as an outgroup a very distantly related species; a too-distant outgroup means a smaller impact on  $d_{AO} - d_{BO}$ . In addition, because the more distantly related the outgroup, the higher the probability that multiple substitutions occurred at some sites, the estimation of the genetic distance is less accurate – even employing a sophisticated model of nucleotide substitution (see Chapter 5). A more powerful test for the molecular clock is the LRT.

### 10.7.2 LRT of the global molecular clock

The phylogeny of a group of taxa is known when the topology and the branch lengths of the phylogenetic tree relating them are known. Of course, whatever the tree topology is, branch lengths can be estimated assuming a constant evolutionary rate along each branch. Clock-like phylogenetic trees are rooted by definition on the longest branch representing the oldest lineage (Figure 10.4). Nonclock-like trees (Figure 10.5A) are unrooted (unless an outgroup is included for rooting the tree; see Chapter 5); in them, a longer branch represents a lineage that evolves faster, which may or may not be an older lineage. Most of the tree-building algorithms, such as the maximum-likelihood, NJ, or *Fitch and Margoliash method*, do not assume a molecular clock; other methods do, such as *UPGMA*.

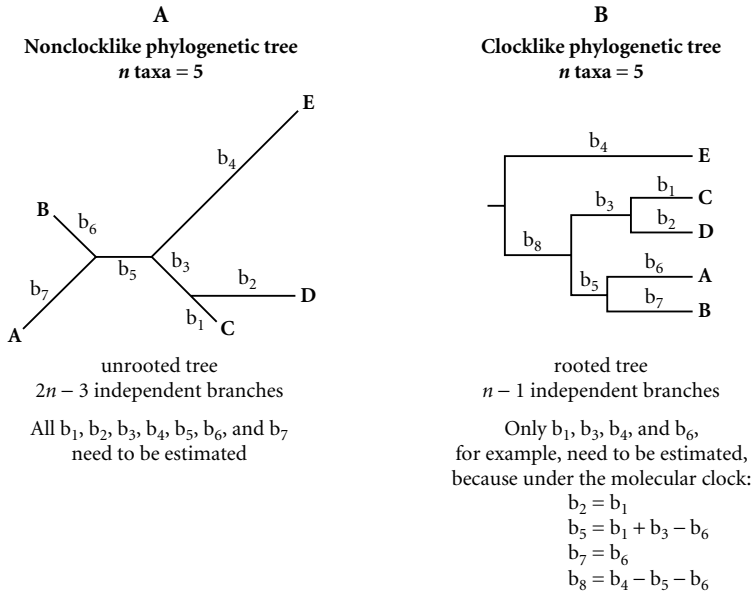


Figure 10.5 Number of free parameters in clock and nonclock trees. Under the free rates model (= nonclock), all the branches need to be estimated ( $2n - 3$ ). Under the molecular clock, only  $n - 1$  branches have to be estimated. The difference in the number of parameters among a nonclock and a clock model is  $n - 2$ .

Maximum-likelihood methods can estimate the branch lengths of a tree by enforcing or not enforcing a molecular clock. In the absence of a molecular clock (the *free-rates model*),  $2n - 3$  branch lengths must be inferred for a strictly bifurcating unrooted phylogenetic tree with  $n$  taxa (Figure 10.5B). If the molecular clock is enforced, the tree is rooted, and just  $n - 1$  branch lengths need to be estimated (see Figure 10.4 and Chapter 1). This should appear obvious considering that under a molecular clock, for any two taxa sharing a common ancestor, only the length of the branch from the ancestor to one of the taxa needs to be estimated, the other one being the same. Statistically speaking, the molecular clock is the null hypothesis (i.e., the rate of evolution is equal for all branches of the tree) and represents a special case of the more general alternative hypothesis that assumes a specific rate for each branch (i.e., free-rates model). Thus, given a tree relating  $n$  taxa, the LRT can be used to evaluate whether the taxa have been evolving at the same rate (Felsenstein, 1988). In practice, a model of nucleotide (or amino-acid) substitution is chosen and the branch lengths of the tree with and without enforcing the molecular clock are estimated. To assess the significance of this test, the LRT can be compared with a  $\chi^2$  distribution with  $(2n - 3) - (n - 1) = n - 2$  degrees of freedom, because the only difference in parameter estimates is in the number of branch lengths that needs to be estimated.

A global molecular clock, ticking at the same rate for all taxa, and a free rate (or nonclock) model, with each taxon evolving at its own rate, are not the only possible scenarios. The clock hypothesis also can be relaxed, allowing a constant rate of evolution within a particular clade but assuming different rates for different clades (i.e., a “local clock” model) (Yoder and Yang 2000). A global molecular clock is a special case of a local molecular clock, which at the same time is a special case of a free-rates model. They can be tested against each other with the LRT; a practical example is discussed later in this chapter. Other relaxation of the molecular clock includes clock models for temporally sampled sequences (i.e., dated tips). These sequences are most frequently from viruses or other fast-evolving pathogens that have been isolated over a range of dates.

An analog to the relative rate test exists within the likelihood framework (Muse and Weir, 1992). Given three nucleotide, amino-acid, or codon sequences and a relevant substitution model, the MLEs can be calculated for the unconstrained three-taxa tree and then for the 3-taxa tree with parameters along two branches constrained to be equal (i.e., the 3rd branch is the outgroup and is estimated independently). A LRT is then performed to determine whether the alternative hypothesis (i.e., all rates are independent) should be accepted or rejected, with the null hypothesis being rates are equal along two given branches.

Molecular-clock calculations based on likelihood methods have been used to date back the origin of viral epidemics, such as the HIV-1 pandemic (Korber et al., 2000; Salemi et al., 2000b), to study substitution dynamics in HIV-1 (Posada and Crandall, 2001b), and to investigate the origin and evolution of the primate T-lymphotropic viruses (PTLVs) (Salemi et al., 1999; Salemi, Desmyter, and Vandamme 2000a). Finally, before applying the LRT for the molecular clock, several precautions need to be taken; specifically, *recombination* has been found to confound this test in such a way that the molecular clock is rejected, when in fact all the lineages are evolving at the same rate (Schierup and Hein, 2000). However, this difficulty can be overcome by using relative ratio tests (Posada, 2001b).

## PRACTICE

David Posada

### 10.8 The model-selection procedure

---

The different model-selection strategies described in the theory section depend on the estimation of likelihood scores, which can be accomplished in programs like PAUP\* (Swofford, 1998), PAML (Yang, 1997), PAL (Drummond and Strimmer, 2001), or HYPHY (Muse and Kosakovsky, 2000). This section demonstrates how to use PAUP\* (see Chapter 7) for selecting models of nucleotide substitution and PAML (Box 10.1) for selecting models of amino-acid replacement.

After the likelihood values of the different candidate models are calculated, the model-selection strategies can be applied easily by hand. In the case of nucleotide substitution models, a user-friendly program called MODELTEST (Posada and Crandall, 1998) facilitates this task. The main steps in the model-selection procedure are as follows:

1. Estimate a tree.
2. Calculate the maximum likelihood of the candidate models, given the data and the tree. This provides the MLEs for the parameters of the model.
3. Compare the likelihood of these models using LRTs or information criteria (i.e., AIC or BIC) to select the best-fit model for the data.

Once a model has been selected, it may be interesting to estimate the parameters of the model (e.g., base frequencies, substitution rates, rate variation) while estimating genetic distances or searching for the best phylogeny, given the model and the data. The user might also want to perform an LRT of the molecular clock using this best-fit model. In fact, the LRT of the molecular clock might be viewed as further model testing; that is, considering the molecular clock as just another parameter that might be added to the model. The first step in the model-selection procedure is the estimation of a tree. In the Theory section of this chapter, this tree was called the base tree. The name comes from the fact that the tree is used only to estimate parameters and likelihoods of different models, rather than being considered the final estimate of the phylogenetic relationships among the taxa under investigation. In fact, it has been shown that as long as this tree is a reasonable estimate of the phylogeny (i.e., a maximum-parsimony or NJ tree; never use a random tree!), the parameter estimates and the model selected will be appropriate. An initial tree can be easily estimated in standard phylogenetic programs like PAUP\* or PHYLIP. Next, the maximum likelihood for each model, given the base tree and the data, needs to be calculated. In practice, the likelihood and the free parameters of the models

**Box 10.1** The PAML package

PAML (Phylogenetic Analysis by Maximum Likelihood) is a freeware software package for phylogenetic analysis of nucleotide and amino-acid sequences using maximum likelihood. Self-extracting archives for MacOS, Windows, and UNIX are available from <http://abacus.gene.ucl.ac.uk/software/paml.html>. The self-extracting archive creates a PAML directory containing several executable applications (extension .exe in Windows or application icons in MacOS), the compiled files (extension .c, placed in the subdirectory src), an extensive documentation (in the doc subdirectory), and several files with example data sets. Each PAML executable also has a corresponding *control file*, with the same name but the extension .ctl, which needs to be edited with a text editor before running the module. For example, the program baseml.exe has a control file called baseml.ctl, which can be opened with any text editor and looks like the following:

```

seqfile = hivALN.phy * sequence data file name
outfile = hivALN.out      * main result file
treefile = hivALN.tre * tree structure filename

noisy = 3 * 0,1,2,3: how much rubbish on the screen
verbose = 0 * 1: detailed output, 0: concise output
runmode = 0 * 0: user tree; 1: semi-automatic; 2: automatic
            * 3: StepwiseAddition; (4,5) :PerturbationNNI

model = 4 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85, 5:TN93,
           6:REV, 7:UNREST
Mgene = 0 * 0:rates, 1:separate; 2:diff pi, 3:diff kapa,
           4:all diff

fix_kappa = 0 * 0: estimate kappa; 1: fix kappa at value below
kappa = 5 * initial or fixed kappa

fix_alpha = 0 * 0: estimate alpha; 1: fix alpha at value below
alpha = 0.3 * initial or fixed alpha, 0:infinity (constant
             rate)

Malpha = 0 * 1: different alpha's for genes, 0: one alpha
ncatG = 8 * # of categories in the dG, AdG, or nparK models
           of rates

clock = 0 * 0:no clock, 1:clock; 2:local clock; 3:TipDate
nhomo = 0 * 0 & 1: homogeneous, 2: kappa for branches, 3:
           N1, 4: N2

getSE = 0 * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 1 * (0,1,2): rates (alpha>0) or ancestral states

```

**Box 10.1** (continued)

```
* Small_Diff = 4e-7
* cleandata = 0 * remove sites with ambiguity data (1:yes, 0:no)?
*   ndata = 1

*   icode = 0 * (with RateAncestor=1. try "GC" in
                data,model=4,Mgene=4)
*   method = 0 * 0: simultaneous; 1: one branch at a time
```

Each executable has a similar control file. The software modules included in PAML usually require an alignment and a tree topology as input. Users have to edit the control file corresponding to the application they want to employ. This editing consists of adding the name of the sequence input file (next to the = sign of the control variable: `seqfile= hivALN.phy` in the previous example), adding the name of the file containing one or more phylogenetic trees for the data set under investigation (next to the = sign of the control variable: `treefile= hivALN.tre` in the previous example), and specifying a name for the output file where results of the computation will be written (`outfile = hivALN.out`). Other control variables are used to choose among different types of analysis. For example, `baseml.exe` can estimate maximum-likelihood parameters of a number of **nucleotide substitution models** (see Chapter 4), given a set of aligned sequences and a tree. The control variable of `baseml.ctl` that needs to be edited in order to choose a model is, in fact, `model`. In the previous example, by assigning `model = 4`, the HKY85 substitution model is chosen (see Section 4.6); most of the other control variables are self-explanatory as well. After editing and saving the control file, the corresponding application (.exe extension) can be executed by simply double-clicking on its icon both in MacOS and Windows. Detailed documentation included in the PAML package (doc subdirectory) should be read before using the software.

**PAML software modules**

The PAML software modules discussed throughout this book are summarized here. Information about the other modules can be found in the PAML documentation.

<b>PAML software module</b>	<b>Input files</b>	<b>Output</b>
<code>baseml.exe</code>	aligned <i>nt</i> sequences, phylogenetic tree The tree also can be estimated by <code>baseml.exe</code> choosing <code>runmode = 2</code> (or 3 or 4) in the control file	ML estimates of different <i>nt</i> substitution models
<code>codeml.exe</code> (see also Section 8.9)	aligned <i>nt</i> coding (or amino-acid) sequences, phylogenetic tree The tree also can be estimated by <code>codeml.exe</code> choosing <code>runmode = 2</code> (or 3 or 4) in the control file.	ML estimates of different amino-acid and nucleotide coding substitution models

`yn00.exe`aligned *nt* coding  
sequencesAnalysis of synonymous  
and nonsynonymous  
replacements in coding  
sequences with the YN98  
method (see Box 11.1  
and Chapter 11)**PAML input files format**

The PAML format is a “relaxed” PHYLIP format (see Box 2.1). Taxa names can be longer than 10 characters and must have at least two blank spaces before starting with the actual sequence. Input trees can be in the usual Newick format (see Figure 5.4). More details can be found in the PAML documentation (in the doc subfolder of the PAML folder).

being compared can be obtained with programs such as PAUP\* or PAML. Once the likelihoods of the different models have been obtained, it is straightforward to apply the LRTs or the AIC procedures. This can be done manually with pencil and paper (and maybe a calculator). Moreover, in the case of the LRTs, a chi-square table is also needed to obtain the p-values. If the number of models compared is high – say, 24 or more models – the model-selection procedure can be tedious. The program MODELTEST (Posada and Crandall, 1998) was designed to help in this task.

**10.9 The program MODELTEST**

MODELTEST is a simple program written in ANSI C and compiled for the Power Macintosh and Windows 95/98/NT using Metrowerks CodeWarrior and for Sun machines using GCC. The MODELTEST package is available for free and can be downloaded from the Web page at [http://bioag.byu.edu/zoology/crandall\\_lab/modeltest.htm](http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm). MODELTEST is designed to compare the likelihood of different nested models of DNA substitution and select the best-fit model for the data set at hand.

The input of MODELTEST is a text file containing a matrix of the log-likelihood scores, corresponding to each one of the 24 nucleotide substitution models shown in Figure 10.1, for a specific data set. Such an input file can be generated by executing a particular block of PAUP\* commands (Box 10.2), which are written in the `modelblock` file included in the MODELTEST package. To test different evolutionary models for a given nucleotide data set, first the sequence input file (in NEXUS format) must be executed in PAUP\* (see Chapter 7). Then, the `modelblock` file can be executed with the data in memory. These commands will make PAUP\* estimate an NJ tree, calculate the likelihood and parameters of

**Box 10.2** PAUP\* command files

Chapter 7 discusses how to use the PAUP\* program by entering commands/options through the command-line interface. Instead of typing all the commands in the command line one by one, separated by a semicolon (see Chapter 7), the user can save them in a text-only document within a so-called PAUP command block, beginning with the keywords `Begin PAUP;` and ending with the keyword `END;` (do not forget the semicolon!). For example, a command block could look like the following:

```
BEGIN PAUP;  
  
Set criterion=distance ;  
  
Dset Distance=JC ;  
  
NJ ;  
  
Lset Rates=gamma Shape=Estimate TRatio=Estimate;  
  
Lscore ;  
  
END ;
```

This file could be saved with the `.nex` extension and successively executed in PAUP\*. Such command files, or batch files, are directly executable in PAUP\* through the `Open . . .` item in the `File` menu. The advantage is that PAUP\* users can write their own scripts to perform complex phylogenetic searches and save them for further analyses. Moreover, such scripts often can be modified easily to perform the same or a similar analysis on different data sets.

the 24 different models, and save the scores to a file called `model.scores`, which will be the input file for `MODELTEST`.

The output of `MODELTEST` consists of a description of the hierarchical LRT and AIC strategies. For hierarchical LRTs, the particular LRTs performed and their associated p-values are listed, and the model selected with the corresponding parameter estimates (actually calculated by PAUP\*) is described. The program also indicates the AIC values and describes the model selected (the one with the smallest AIC) with the corresponding parameter estimates. The output of `MODELTEST` also provides a block of commands in `NEXUS` format, which can be executed in PAUP\* with the sequence data in memory to automatically implement the selected model. This is useful if the user wants to implement the selected model in PAUP\* for further analysis (e.g., to perform an LRT of the molecular clock or to estimate a phylogenetic tree using the best-fit model).

In summary, testing nucleotide substitution models with `MODELTEST` consists of the following steps:

1. Open the data file and execute it in PAUP\*.
2. Execute the command file `modelblock3` located in the `MODELTEST` folder. PAUP\* estimates an NJ tree and the likelihood and parameter values for several

models. The task can take from several minutes to several hours, depending on the number of taxa and the computer speed. Once finished, a file called `model.scores` will appear in the same directory as the `modelblock` file.

3. Execute MODELTEST with the file `model.scores`, output from the previous step, as input file. The Mac version of the program has a command-line interface asking the user to select an input file and choose a name for the output file. The PC version requires `model.scores` to be in the same directory where `modeltest.exe` is (this directory, called `Modeltest`, is created during installation of the program). When executing the program, an MS-DOS window appears. To implement the computation, type `modeltest.exe < model.scores > outfile` and press enter. The program will save the `outfile` with results in the same directory.

### 10.10 Implementing the LRT of the molecular clock using PAUP\*

Once a substitution model has been selected by MODELTEST, the LRT of the molecular clock can be performed using the current likelihood of the model and a new likelihood can be calculated enforcing a molecular clock on the tree. As discussed in the previous section, the execution of `modelblock3` makes PAUP\* infer a simple NJ tree with Jukes and Cantor distances, and uses the tree to estimate likelihood and parameters of the other evolutionary models as well. Therefore, it is possible to evaluate the clock hypothesis by calculating the likelihood of the rooted version of this tree enforcing a molecular clock. The likelihood of such a tree can be compared in an LRT with the likelihood obtained for the correspondent nonclock model, which can be found in the MODELTEST output file. The calculation can be implemented in PAUP\* as follows:

1. Infer the NJ tree in PAUP\* by executing the *PAUP command block* (see Section 7.8) as follows:

```
BEGIN PAUP;

DSet distance=JC objective=ME base=equal rates=equal
pinv=0 subst=all negbrlen=setzero;

NJ showtree=no breakties=random;

END;
```

This is precisely the first command block of `modelblock3`. It computes a simple NJ tree with distances estimated with the Jukes and Cantor model.

2. The tree has to be rooted to implement the clock parametrization. This can be achieved with the `root` command, either by choosing an outgroup, if available, or by midpoint rooting.
3. As discussed in the previous section, the output file of MODELTEST contains a command block specifying the parameters of the selected model. Add the PAUP\*

command `clock=yes` to the end of the `Lset` block, before the semicolon, and save the entire command block in a separate document as text-only using any text editor. Eventually, the command block will be something like the following:

```
BEGIN PAUP;

Lset Base=(0.4159 0.2281 0.1269) Nst=6 Rmat=(1.0000
2.8596 1.0000 1.0000 5.7951) Rates=gamma Shape=0.6806
Pinvar=0.1698 clock=yes;

END;
```

This PAUP\* command block, for example, provides the likelihood settings for the TN +  $\Gamma$  + I model. It specifies the *relative rate parameters* of the distance matrix, the shape parameter of the  $\Gamma$ -*distribution* ( $\alpha = 0.6806$  in this case), and the proportion of invariable sites (Pinvar = 0.1698), which have all been estimated by PAUP\* when executing `modelblock3` (see Chapter 7) using the same NJ tree in memory.

4. Execute the command block in PAUP\* (see Chapter 7). The program estimates the log likelihood of the model under the molecular clock, with  $L_0$  representing the probability of the null hypothesis. The log likelihood of the model not enforcing the clock,  $L_1$ , is the log likelihood of the selected model written in the output file of MODELTEST.
5. The LRT can now be done manually. Calculate  $\Delta = 2 * (L_1 - L_0)$ . Because both values  $L_0$  and  $L_1$  are negative, but being that  $L_1$  is bigger than  $L_0$ , the  $\Delta$  value should be positive. The number of degrees of freedom will be the number of taxa  $- 2$  (see Section 10.7). The corresponding p-value can be found in a chi-square table. Alternatively, MODELTEST can be used to implement the LRT. Execute the program with the option `-c` in the argument line (see documentation for different operating systems). Input  $|L_0|$  (the absolute value of  $L_0$ ),  $|L_1|$  (the absolute value of  $L_1$ ), and the number of degrees of freedom (number of taxa  $- 2$ ).

The p-value is interpreted as the probability of observing the obtained LRT statistic ( $\Delta$ ) if the taxa are evolving according to a molecular clock. In other words, if this value is smaller than 0.05 (or 0.01, if a less conservative test is preferred), the molecular clock hypothesis is rejected. When the p-value is marginally significant (close to 0.10–0.01), a more strict way of performing the LRT test would be to use a maximum-likelihood tree. In such a case, first estimate the ML tree with the best-fitting model – which also gives the likelihood of the model without assuming a clock – and then estimate the likelihood of the same model enforcing the clock on the tree.

### 10.11 Selecting the best-fit model in the example data sets

The first two example data sets were analyzed as described previously using MODELTEST and PAUP\*. The candidate models compared were JC, JC + I,

**Table 10.1A** Hierarchical LRT of models of molecular evolution for the mtDNA data

Null hypothesis	Models compared	-ln Likelihoods	LRT 2(ln L <sub>1</sub> - ln L <sub>0</sub> )	df	P-value
Equal base frequencies	H <sub>0</sub> : JC69	-ln L <sub>0</sub> : 23646	420	3	<0.000001
	H <sub>1</sub> : F81	-ln L <sub>1</sub> : 23436			
Equal ti/tv rates	H <sub>0</sub> : F81	-ln L <sub>0</sub> : 23436	854	1	<0.000001
	H <sub>1</sub> : HKY	-ln L <sub>1</sub> : 23009			
Equal ti and equal tv rates	H <sub>0</sub> : HKY	-ln L <sub>0</sub> : 23009	664	4	<0.000001
	H <sub>1</sub> : GTR	-ln L <sub>1</sub> : 22677			
Equal rates among sites	H <sub>0</sub> : GTR	-ln L <sub>0</sub> : 22677	3042	1	<0.000001
	H <sub>1</sub> : GTR + Γ	-ln L <sub>1</sub> : 21156			
Proportion of invariable sites	H <sub>0</sub> : GTR + Γ	-ln L <sub>0</sub> : 21156	16	1	0.000063
	H <sub>1</sub> : GTR + Γ + I	-ln L <sub>1</sub> : 21148			
Molecular clock	H <sub>0</sub> : GTR + Γ + I/clock	-ln L <sub>0</sub> : 21192	88	15	<0.000001
	H <sub>1</sub> : GTR + Γ + I	-ln L <sub>1</sub> : 21148			

JC + I + Γ, K80, K80 + I, K80 + I + Γ, SYM, SYM + I, SYM + I + Γ, F81, F81 + I, F81 + I + Γ, HKY, HKY + I, HKY + I + Γ, GTR, GTR + I, and GTR + I + Γ, where I means there is a significant proportion of invariable sites, and Γ means a gamma distribution is being used to account for rate variation among sites.

**10.11.1 Vertebrate mtDNA**

The first data set is an alignment of mitochondrial sequences from several vertebrates. The model selected by the hierarchical LRTs, dynamical LRTs, AIC, and BIC was the GTR + I + Γ model (Table 10.1A), which is the most complex of the candidate models. In this model, base frequencies are unequal, the six possible types of substitutions among the different bases occur at different rates, there is a significant proportion of invariable sites (p-inv), and some sites evolve faster than others because α, the shape parameter of the gamma distribution, is smaller than 1 (Table 10.1B). The estimates obtained characterize the molecular evolution of this gene: A is the most frequent nucleotide, the most common substitution is between C and T, around 16% of the sites are invariable, and there is medium-rate heterogeneity. As explained in the Theory section, for an idea of how much this model describes the data, it is possible to compare its likelihood to the likelihood of the multinomial model. The log of the likelihood rather than the likelihood itself is

**Table 10.1B** Estimated parameters for the GTR + Γ + I model

nt Frequencies		Relative substitution rates (G ↔ T = 1.0)									
-lnL	π <sub>A</sub>	π <sub>C</sub>	π <sub>G</sub>	π <sub>T</sub>	A ↔ C	A ↔ G	A ↔ T	C ↔ G	C ↔ T	p-inv	α
21148	0.38	0.22	0.18	0.22	3.41	5.10	3.51	0.44	14.99	0.16	0.73

**Table 10.2** Hierarchical LRT of models of molecular evolution for the HIV *env* data

Null hypothesis	Models compared	–ln Likelihoods	LRT 2(ln L <sub>1</sub> – ln L <sub>0</sub> )	df	P-value
Equal base frequencies	H <sub>0</sub> : JC69	–ln L <sub>0</sub> : 22100	366	3	<0.000001
	H <sub>1</sub> : F81	–ln L <sub>1</sub> : 21917			
Equal ti/tv rates	H <sub>0</sub> : F81	–ln L <sub>0</sub> : 21917	748	1	<0.000001
	H <sub>1</sub> : HKY	–ln L <sub>1</sub> : 21543			
Equal ti and equal tv rates	H <sub>0</sub> : HKY	–ln L <sub>0</sub> : 21543	154	4	<0.000001
	H <sub>1</sub> : GTR	–ln L <sub>1</sub> : 21466			
Equal rates among sites	H <sub>0</sub> : GTR	–ln L <sub>0</sub> : 21466	1638	1	<0.000001
	H <sub>1</sub> : GTR + $\Gamma$	–ln L <sub>1</sub> : 20647			
Proportion of invariable sites	H <sub>0</sub> : GTR + $\Gamma$	–ln L <sub>0</sub> : 20647	10	1	0.001565
	H <sub>1</sub> : GTR + $\Gamma$ + I	–ln L <sub>1</sub> : 20642			
Molecular clock	H <sub>0</sub> : GTR + $\Gamma$ + Ic	–ln L <sub>0</sub> : 20653	22	12	0.037520
	H <sub>1</sub> : GTR + $\Gamma$ + I	–ln L <sub>1</sub> : 20642			

compared, because likelihoods are often too small to be handled appropriately by standard computers.

To estimate the likelihood of the multinomial model in PAUP\*, it is necessary to eliminate positions with ambiguities from the data. The log likelihoods of the GTR + I +  $\Gamma$  model and the multinomial model are then, respectively, –20810 and –11099, which indicates that the selected model does not explain much of the data.

The LRT of the molecular clock can be performed as described previously. The log likelihood of the GTR + I +  $\Gamma$  model under the molecular clock is –21192, which is significantly smaller than the likelihood without assuming a clock (see Table 10.1). Consequently, the global molecular clock hypothesis is rejected for this data set.

### 10.11.2 HIV envelope gene

The second data set is the *env* gene from HIV and SIV. Again, the model selected by the hierarchical LRTs, dynamical LRTs, AIC, and BIC was the GTR + I +  $\Gamma$  model (Table 10.2). A is the most frequent nucleotide, the most common substitution is between A and G, approximately 13% of the sites have not changed, and there is medium-rate heterogeneity.

Again, to estimate the likelihood of the multinomial model in PAUP\*, ambiguities were eliminated from the data. After doing that, the log likelihoods of the GTR + I +  $\Gamma$  model and the multinomial model are, respectively, –15080 and –11361, which indicates that the selected model reasonably explains the data.

The log likelihood of the GTR + I +  $\Gamma$  model under the molecular clock is –20653, which is significantly smaller than the likelihood without assuming a clock (see Table 10.2). However, the p-value is close to being nonsignificant;

**Table 10.3** AIC values for different models of amino-acid replacement in the enzyme glycerol-3-phosphate dehydrogenase in bacteria

Model <sup>1</sup>	−ln L	$\alpha^2$	Free parameters	AIC
Poisson	7704	$\infty$	0	15408
Proportional	7533	$\infty$	19	15104
<i>Empirical</i>				
Jones	7202	$\infty$	0	14404
Dayhoff	7246	$\infty$	0	14492
WAG	7117	$\infty$	0	14234
<i>Empirical + F</i>				
Jones	7208	$\infty$	19	14454
Dayhoff	7246	$\infty$	19	14530
WAG	7110	$\infty$	19	14258
REVAAL_0	7205	$\infty$	93	14596
Poisson + $\Gamma$	7650	2.5	1	15302
Proportional + $\Gamma$	7476	2.3	20	14992
<i>Empirical + <math>\Gamma</math></i>				
Jones	7094	1.66	1	14190
Dayhoff	7125	1.56	1	14252
WAG	7043	2.13	1	<b>14088</b>
<i>Empirical + F + <math>\Gamma</math></i>				
Jones	7099	1.66	20	14238
Dayhoff	7124	1.54	20	14288
WAG	7037	2.13	20	14114
REVAAL_0 + $\Gamma$	7076	0.004	94	14340

Note: Likelihood values were estimated in PAML 3.0b (Yang, 1997). The model with smallest AIC value is in boldface.

<sup>1</sup> Poisson (Zuckerandl and Pauling, 1965), Proportional (Hasegawa and Fujiwara, 1993), Jones (Jones et al., 1992), Dayhoff (Dayhoff et al., 1978; Kishino et al., 1990), WAG (Whelan and Goldman, in press), REVAAL\_0 (Yang et al., 1998); + F: including amino-acid frequencies observed from the data; +  $\Gamma$ : including rate variation as described by the gamma distribution.

<sup>2</sup>  $\alpha$  is the shape parameter of the gamma distribution.

therefore, the conclusion is not definitive. A larger and more representative HIV data set would be needed to address the issue.

### 10.11.3 G3PDH protein

The third data set is an amino-acid alignment of the enzyme glycerol-3-phosphate dehydrogenase in bacteria, protozoa, and animals. Because not all models compared (see Chapter 8) are nested, the AIC criterion was used in this case. The model with the best AIC values was the empirical model with the WAG amino-acid replacement

matrix (Whelan and Goldman, 2001) with rate variation among sites ( $WAG + \Gamma$ ). The estimated value of the shape parameter of the gamma distribution was 2.13, which indicates that there is moderate rate variation among sites. To estimate the likelihood of the multinomial model in PAML, ambiguities were eliminated from the data. After removing ambiguous positions, the log likelihoods of the WAG model and the multinomial model are, respectively,  $-5523$  and  $-1574$ , which indicates that the selected model inadequately explains the data. The log likelihood of the  $WAG + \Gamma$  model under the molecular clock is  $-7119$ , which is significantly smaller than the likelihood without assuming a clock (Table 10.3). Consequently, the molecular clock hypothesis should be rejected.

## REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt (1978). A model of evolutionary change in proteins. In: *Atlas of Protein Sequence and Structure*, ed. M. O. Dayhoff, pp. 345–352. Washington, DC.
- Drummond, A. and K. Strimmer (2001). PAL: An object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, 17, 662–663.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum-likelihood approach. *Journal of Molecular Evolution*, 17, 368–376.
- Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Reviews in Genetics*, 22, 521–565.
- Goldman, N. (1993a). Simple diagnostic statistical test of models of DNA substitution. *Journal of Molecular Evolution*, 37, 650–661.
- Goldman, N. (1993b). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, 36, 182–198.
- Hartl, D. L. and A. G. Clark (1997). *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates, Inc.
- Hasegawa, M. and M. Fujiwara (1993). Relative efficiencies of the maximum-likelihood, maximum-parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution*, 2, 1–5.
- Hasegawa, M., K. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22, 160–174.
- Hillis, D. M., C. Moritz, and B. K. Mable (1996). *Molecular Systematics*. Sunderland, MA: Sinauer Associates, p. 655.
- Huelsenbeck, J. P. and K. A. Crandall (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Reviews in Ecological Systems*, 28, 437–466.
- Huelsenbeck, J. P. and B. Rannala (1997). Phylogenetic methods come of age: Testing hypothesis in an evolutionary context. *Science*, 276, 227–232.

- Huelsenbeck, J. P., D. M. Hillis, and R. Jones (1996). Parametric bootstrapping in molecular phylogenetics: Applications and performance. In: *Molecular Zoology: Advances, Strategies, and Protocols*, eds. J. D. Ferraris and S. R. Palumbi, pp. 19–45. New York: Wiley-Liss.
- Jones, D. T., W. R. Taylor, and J. M. Thornton (1992). The rapid generation of mutation data matrixes from protein sequences. *Computer Applications in the Biosciences*, 8, 275–282.
- Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In: ed. H. M. Munro, *Mammalian Protein Metabolism*, pp. 21–132. New York: Academic Press.
- Kass, R. E. and L. Wasserman (1994). *A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion*. Pittsburgh, PA: Carnegie Mellon University, Department of Statistics, p. 16.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217, 624–626.
- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111–120.
- Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Kishino, H., T. Miyata, and M. Hasegawa (1990). Maximum likelihood inferences of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31, 151–160.
- Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288, 1789–1796.
- Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.
- Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA: Sinauer Associates.
- Liò, P. and N. Goldman (1998). Models of molecular evolution and phylogeny. *Genome Research*, 8, 1233–1244.
- Muse, S. V. and S. L. Kosakovsky (2000). *HYPHY: Hypothesis testing using phylogenies*. Raleigh, NC: North Carolina State University, Department of Statistics, Program in Statistical Genetics.
- Muse, S. V. and B. S. Weir (1992). Testing for equality of evolutionary rates. *Genetics*, 132, 269–276.
- Page, R. D. M. and E. C. Holmes (1998). *Molecular Evolution: A Phylogenetic Approach*. Abingdon, UK: Blackwell Science.
- Posada, D. (2001a). The effect of branch-length variation on the selection of models of molecular evolution. *Journal of Molecular Evolution*, 52, 434–444.
- Posada, D. (2001b). Unveiling the molecular clock in the presence of recombination. *Molecular Biology and Evolution*, 18, 1976–1978.
- Posada, D. and K. A. Crandall (1998). Modeltest: Testing the model of DNA substitution. *Bioinformatics*, 14, 817–818.
- Posada, D. and K. A. Crandall (2001a). Selecting the best-fit model of nucleotide substitution. *Systematics Biology*, 50, 580–601.
- Posada, D. and K. A. Crandall (2001b). Selecting models of nucleotide substitution: An application to the human immunodeficiency virus 1 (HIV-1). *Molecular Biology and Evolution*, 18, 897–906.

- Rodríguez, F., J. F. Oliver, A. Marín, and J. R. Medina (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142, 485–501.
- Rzhetsky, A. and M. Nei (1995). Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution*, 12, 131–151.
- Salemi, M., M. J. Lewis, J. F. Egan, W. W. Hall, J. Desmyter, and A.-M. Vandamme (1999). Different population dynamics and evolutionary rates of human T-cell lymphotropic virus type II (HTLV-II) in injecting drug users compared to in endemically infected Amerindian and Pygmy tribes. *Proceedings of the National Academy of Sciences of the USA*, 96, 13253–13259.
- Salemi, M., J. Desmyter, and A.-M. Vandamme (2000a). Evolutionary history of human and simian T-lymphotropic viruses (HTLV/STLVs) revealed by analyses of full genome sequences. *Molecular Biology and Evolution*, 3, 374–386.
- Salemi, M., K. Strimmer, W. W. Hall, M. Duffy, E. Delaporte, S. Mboup, M. Peeters, and A.-M. Vandamme (2000b). Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clocklike molecular evolution. *The FASEB Journal*, 2001:15:276–278.
- Schierup, M. H. and J. Hein (2000). Recombination and the molecular clock. *Molecular Biology and Evolution*, 17, 1578–1579.
- Schwarz, G. (1974). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Swofford, D. L. (1998). *PAUP\* Phylogenetic analysis using parsimony and other methods*. Sunderland, MA: Sinauer Associates.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis (1996). Phylogenetic inference. In: *Molecular Systematics*, eds. D. M. Hillis, C. Moritz, and B. K. Mable, pp. 407–514. Sinauer Associates, Sunderland, MA.
- Whelan, S. and N. Goldman (1999). Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, 16, 1292–1299.
- Whelan, S. and N. Goldman (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Molecular Biology and Evolution*, 18, 691–699.
- Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple-sequence data. *Journal of Molecular Evolution*, 42, 587–596.
- Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13, 555–556.
- Yang, Z., R. Nielsen, and H. Masami (1998). Models of amino-acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15, 1600–1611.
- Yoder A. D. and Z. Yang (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17, 1081–1090.
- Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39, 315–329.
- Zuckerandl, E. and L. Pauling (1962). Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in Biochemistry*, eds. M. Kasha and B. Pullman, pp. 189–225. Academic Press, New York.
- Zuckerandl, E. and L. Pauling (1965). Evolutionary divergence and convergence in proteins. In: *Evolving Genes and Proteins*, eds. V. Bryson and H. J. Vogel, pp. 97–166. Academic Press, New York.