Detecting the Signatures of Adaptive Evolution in Protein-Coding Genes

Joseph P. Bielawski[1]

[1]Department of Biology, Department of Mathematics & Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

ABSTRACT

The field of molecular evolution, which includes genome evolution, is devoted to finding variation within and between groups of organisms and explaining the processes responsible for generating this variation. Many DNA changes are believed to have little to no functional effect, and a neutral process will best explain their evolution. Thus, a central task is to discover which changes had positive fitness consequences and were subject to Darwinian natural selection during the course of evolution. Due the size and complexity of modern molecular datasets, the field has come to rely extensively on statistical modeling techniques to meet this analytical challenge. For DNA sequences that encode proteins, one of the most powerful approaches is to employ a statistical model of codon evolution. This unit provides a general introduction to the practice of modeling codon evolution using the statistical framework of maximum likelihood. Four real-data analysis activities are used to illustrate the principles of parameter estimation, robustness, hypothesis testing, and site classification. Each activity includes an explicit analytical protocol based on programs provided by the Phylogenetic Analysis by Maximum Likelihood (PAML) package. *Curr. Protoc. Mol. Biol.* 101:19.1.1-19.1.21. © 2013 by John Wiley & Sons, Inc.

Keywords: molecular evolution • protein evolution • selection pressure • codon models • maximum likelihood

INTRODUCTION

Gene and genome evolution is most often studied in terms of the outcome of the process, i.e., to infer the pattern of phylogenetic relationships among sequences. However, a sample of gene sequences can be used to investigate the evolutionary process itself. Statistical modeling techniques provide a powerful framework for this; however, they are often computationally intensive. Rapid increases in computer performance over the last decade, coupled with decreases in cost, are leading to the use of statistical modeling techniques in almost every aspect of molecular evolution and genomics. This unit focuses on modeling the process of codon evolution for the purpose of investigating the role of natural selection during gene sequence evolution. It contains separate modeling activities (each with a protocol) dedicated to (1) parameter estimation, (2) robustness, (3) modeling changes in selection pressure over time, and (4) modeling variable selection pressure among sites within a gene. Collectively, these activities are designed to introduce biologists to the statistical techniques required to analyze protein-coding genes and to detect the signature of molecular adaption. The unit is suitable for people with no prior experience with codon models, but it does assume some familiarity with sequence alignment and phylogenetic analysis.

Brief introduction to codon models

Codon models specify the process of evolving from one codon state to another within a protein-coding sequence of DNA. There are many advantages to modeling at the codon level, as compared to using nucleotide or amino acid states; chief among them is the

Current Protocols in Molecular Biology 19.1.1-19.1.21, January 2013
Published online January 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142727.mb1901s101

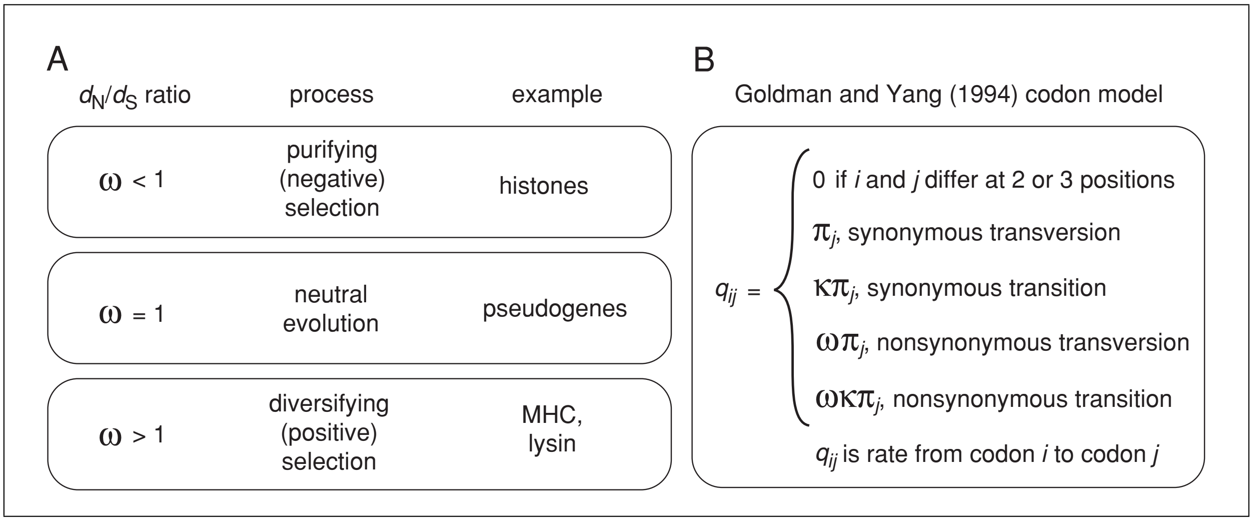

A
d_N/d_S ratio
process
example
ω < 1
purifying (negative) selection
histones
ω = 1
neutral evolution
pseudogenes
ω > 1
diversifying (positive) selection
MHC, lysin
B
Goldman and Yang (1994) codon model
q_{ij} =
0 if i and j differ at 2 or 3 positions
π_j, synonymous transversion
$\kappa\pi_j$, synonymous transition
$\omega\pi_j$, nonsynonymous transversion
$\omega\kappa\pi_j$, nonsynonymous transition
q_{ij} is rate from codon i to codon j

Figure 19.1.1 Modeling the intensity of natural selection via a codon model. (**A**) The ω ratio (d_N/d_S) is a parameter used to measure the direction and intensity of natural selection pressure acting on a protein. (**B**) Codon models specify the probability of substitution between the sense codons within a protein sequence, which depends on the value of the ω parameter. Given an explicit codon model, such as Goldman and Yang (1994), the value of the ω parameter can be estimated from a dataset via the method of maximum likelihood. Other model parameters are the transition/transversion ratio (κ) and the frequency of the j^{th} codon (π_j). MHC, major histocompatibility complex.

capacity to distinguish between the synonymous rate (d_S) and the nonsynonymous rate (d_N) of evolution within a gene. The ratio of these rates (d_N/d_S), hereafter referred to as ω, has proven to be a useful measure of the direction and intensity of natural selection pressure acting on a protein (Yang and Bielawski, 2000; Anisimova and Liberles, 2012; Fig. 19.1.1A). For example, if selection were not acting (i.e., neutral evolution), the rate of nonsynonymous evolution would be the same as the synonymous rate, with $\omega = 1$. However, most proteins are dominated by purifying selection (i.e., the removal of functionally deleterious mutations) and their nonsynonymous rate will be less than the synonymous rate, with $\omega < 1$. Stronger purifying selection pressure leads to lower values of ω. Rarely, a protein will experience a burst of changes, driven by selection, that increase fitness (i.e., positive, or Darwinian, selection). In such cases the nonsynonymous rate can exceed the synonymous rate, with $\omega > 1$. More intense positive selection leads to higher values of ω. Because many evolutionary biologists utilize codon models as a tool to investigate the process of molecular innovation and divergence, codon models have been the subject of substantial research efforts.

Goldman and Yang (1994) and Muse and Gaut (1994) independently proposed similar codon models, and these serve as the foundation for most of those presently in use. They model codon evolution as a Markov process; that is, they specify a simple process where the probability of a substitution of one codon to another depends only on the current state of the codon and not on any past codon states (Fig. 19.1.1B). This Markov process describes the substitutions between the sense codons within a protein sequence (e.g., there are 61 sense codons in the universal genetic code). Substitutions that produce stop codons are not included in the model because they are not tolerated within a functional protein-coding gene. For practical purposes, evolution is usually assumed to be independent among sites. Hence, at a given site the model is used to specify the instantaneous rate of substitution from a particular codon i to another codon j, and this is denoted q_{ij}. The parameter ω is employed to account for the effect of selection acting on the protein product of the gene; if a substitution is nonsynonymous, its rate in the model is multiplied by ω (Fig. 19.1.1B). Hence, purifying selection ($\omega < 1$) reduces the nonsynonymous rate relative to the synonymous rate of evolution in that gene. Other evolutionary processes also can be explicitly modeled. For example, DNA transitions (i.e., A ↔ G or T ↔

C) often evolve faster than DNA transversions. In the model of Goldman and Yang (1994), all substitutions between codons that involve a DNA transition are multiplied by a transition/transversion rate ratio (κ).

Codon models have been extended in a bewildering variety of ways. They have been modified to permit selection pressures (ω) to vary over time (e.g., Muse and Gaut, 1994; Yang, 1998), over sites (e.g., Yang et al., 2000; Yang and Swanson, 2002), and both (e.g., Yang and Nielsen, 2002; Bielawski and Yang, 2004). Some models now permit variability among sites in other aspects of evolution, such as the transition/transversion ratio, codon frequencies, and synonymous rates (Bao et al., 2007, 2008). Others attempt to model non-independence among sites, and still others to model a covarion-like substitution process among codons (Guindon et al., 2004). The details of these, and many other models, are beyond the scope of this work. The reader is referred to Bielawski and Yang (2005) for further information about models permitting selection pressures to vary over time and over sites. Readers interested in a broad survey of models and methodological developments are referred to Anisimova and Kosiol (2009).

Maximum likelihood estimation of model parameters

The likelihood framework provides us with a means of making inferences about the process that generated the data we have "in hand." In this case we have in hand a set of DNA sequences, and our task is to identify the evolutionary processes that provide the best explanation of those data. Here, it becomes worthwhile to make a distinction between the concepts of probability and likelihood. Probability refers to the degree to which some future event is certain or uncertain. This notion is often illustrated with coin flipping; for example, if you had a fair coin and you planned to toss it 5 times, you could easily compute the probability that you would observe 4 heads and 1 tail (0.1563 via the binomial). There are two preconditions associated with this example: (1) the outcome is a *variable* because it has not yet been observed, and (2) the hypothesis that the coin is indeed "fair" is a *fixed quantity*. Likelihood differs by exchanging what is variable and what is fixed. Under likelihood, the first precondition is that the outcome has already been observed and thus it is a *fixed quantity* (say, the coin was flipped 5 times and precisely 3 heads and 2 tails were observed; we must now deal with these data rather than some hypothetical outcome). The second precondition is that truth of the hypothesis of "fairness" is unknown, so it is treated as a *variable*. Thus we can ask the question: *What is the likelihood that my coin is fair given that I have observed 3 heads and 2 tails*? The central point is that likelihood is often employed when the task is to learn which of several hypotheses provides the best explanation for the data in hand. A more formal introduction to likelihood is not practical here, and interested readers are referred to Pawitan (2001).

Since we always work with a precise set of sequences [i.e., our multiple sequence alignment (MSA)], the dataset must be treated as a fixed quantity. On the other hand, the intensity of natural selection during the evolution of those sequences is an unknown variable. So, the likelihood framework is a natural way to approach the question: *What is the likelihood that the sequences in my dataset evolved under positive selection*? For parameter estimation, a codon model is used to compute the likelihood of the sequence data under different values for selection pressure (ω), as well as any other parameters concerning the evolutionary process. The values of the model parameters that maximize the likelihood of the sequence data are taken as the best estimates of the parameters. Through these maximum likelihood estimates (MLEs) we can learn about the evolutionary processes that generated the observed sample of sequences.

This concept is illustrated with a simple but real sequence dataset. The data are the *GstD1* gene sequences (600 codons) from *Drosophila melanogaster* and *Drosophila simulans*.

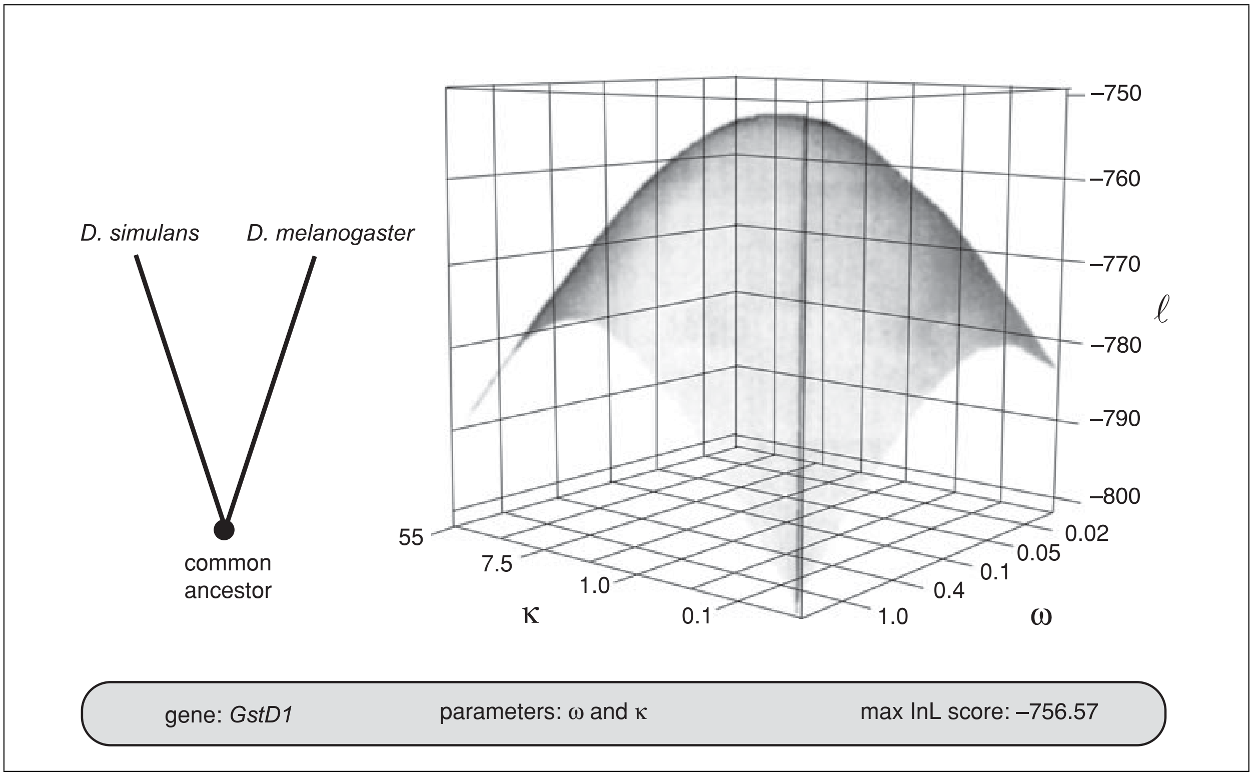


Figure 19.1.2 Likelihood of the *GstD1* sequences from *Drosophila simulans* and *Drosophila melanogaster* as a function of both the ω and κ parameters of a codon model. The values of ω and κ that maximize the likelihood of the data (–756.57) are 0.067 and 2.53 respectively.

A simple codon model having parameters for selection pressure (ω), transition bias (κ) and overall sequence divergence (t) was employed to compute the likelihood (L) of the data for a wide variety of parameter values. Because the calculations involve very small numbers, the likelihood scores were transformed to log likelihoods (ℓ). Figure 19.1.2 shows the relationship between the log likelihood of the data and values for parameters ω and κ of the codon model. Note that there is a "surface" to the log-likelihood scores with a peak that corresponds to $\omega = 0.067$ and $\kappa = 2.53$; these are the MLEs for this dataset. To be a little more formal about this: the MLEs are the values that maximize the likelihood function and are denoted $\hat{\omega}$ and $\hat{\kappa}$. With $\hat{\omega}$ (0.067) being considerably less than 1, these sequences seem to have been dominated by purifying selection pressure. Also, $\hat{\kappa}$ (2.53) suggests that transitions have occurred at a higher rate than transversions. The protocol for Activity 1 is designed to illustrate the concept of parameter estimation in a very simple setting.

Assessing the model

Evident from Figure 19.1.2 is that the likelihood of our data depends on the parameters of the model. Since the ultimate goal is to draw conclusions about the evolutionary processes that generated the data, it is often desirable to assess the reliability of the parameter estimates. There are many different ways to model codon evolution, so let us define θ as the parameters we intend to estimate from whatever model is chosen. The parameter values that maximize the likelihood of the data will be denoted $\hat{\theta}$. The shape of the likelihood surface around $\hat{\theta}$ contains information about the reliability of those estimates. Specifically, the steeper the curve in likelihood around $\hat{\theta}$, the more informative the data are about the values of $\hat{\theta}$. One way to quantify how certain we can be about the values of $\hat{\theta}$ is to identify a region centered on the peak that contains, say, a 95% likelihood region for θ. This is based on the likelihood ratio (see next section), and is defined by those values of θ for which the likelihood ratio is less than the critical percentile of χ_k^2, where k is the number of free model parameters [see Yang (2006) for additional details and examples]. The approach assumes that the dataset is large enough, and the model is not too rich, so that the asymptotic distribution is applicable. This will often not be the case and χ_k^2 will only be an approximation. Nonetheless, the approach is useful; it is common to observe that some parameters are easier to estimate than others (i.e., the likelihood drops more steeply for some parameters than for others with increasing distance from their MLEs). Users should note that the parameters of codon models are difficult to estimate, often having large amounts of uncertainty.

Another aspect of model reliability is the sensitivity of results to assumptions about the evolutionary process that are included in the model. Often, a user is interested in just one parameter, say ω, with the other model parameters being so-called nuisance parameters. The nuisance parameters are nonetheless biologically relevant (e.g., the transition/transversion ratio, κ) and how they are treated in the model can be very important for some datasets (Bao et al., 2007). Inappropriate modeling can sometimes lead to false biological conclusions (Bao et al., 2008). The protocol for Activity 2 presents a robustness analysis on a small dataset, and it illustrates how estimates of the intensity of selection pressure can be very sensitive to assumptions about the transition/transversion ratio and codon frequencies.

Hypothesis testing under maximum likelihood

Parameter estimation is only one aspect of a maximum likelihood analysis. The likelihood framework also provides a powerful method for testing hypotheses. The method is called the likelihood ratio test (LRT), and it combines parameter estimation with a general method for comparing two models that embody two competing hypotheses. If H_0 and H_1 indicate competing hypotheses, then let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the MLEs under the models for those hypotheses. The ratio of likelihood scores under $\hat{\theta}_0$ and $\hat{\theta}_1$ is the test statistic

$$2\Delta\ell = 2\log\left(\frac{L_1(\hat{\theta}_1)}{L_0(\hat{\theta}_0)}\right) = 2\left(\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)\right)$$

When H_0 is true, $2\Delta\ell$ is approximately χ^2 distributed with degrees of freedom equal to the difference in the number of parameters between the two models. As already mentioned, the dataset must be large enough for asymptotic distribution to be applicable. Furthermore, the models must be nested (i.e., H_0 is a constrained version of H_1) and H_0 should not be a case of H_1 with one or more of its parameters fixed on the boundary of the parameter space. Monte Carlo simulation can be employed to obtain the correct distribution when the above conditions have not been met.

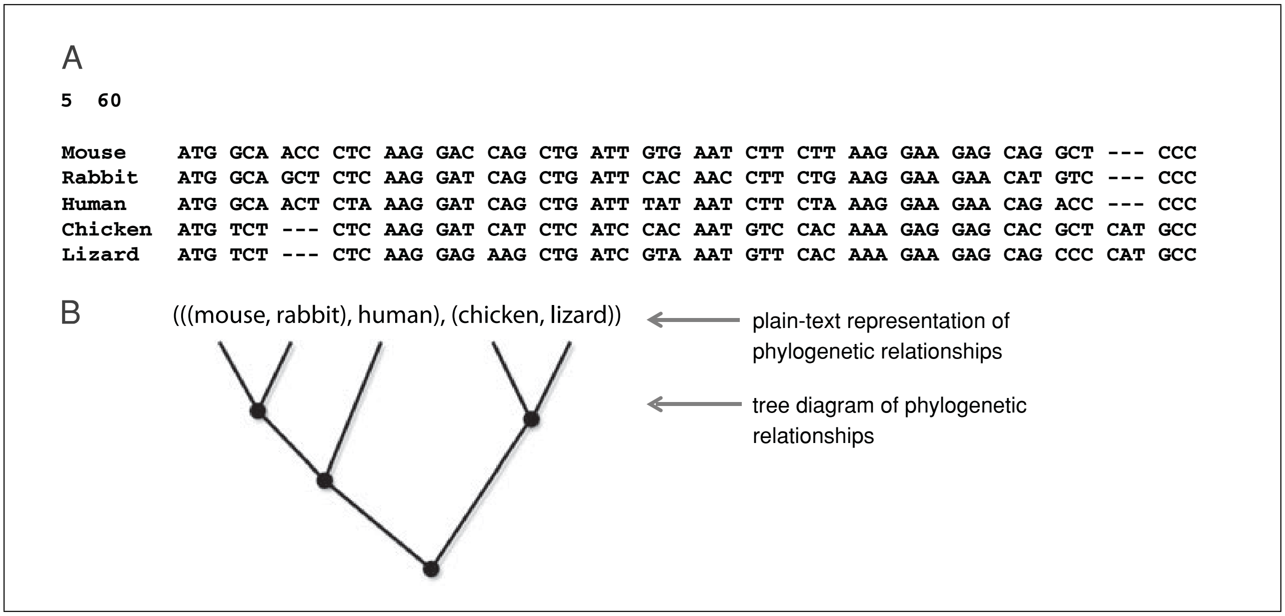
LRTs based on codon models have been employed to test a wide variety of hypotheses about the process of protein evolution (Bielawski and Yang, 2005). Among the most widely used are (1) tests for variation in selective pressures among branches and (2) tests for variable selective pressures among sites. A subset of the latter type of test can be formulated as an explicit test for the action of positive Darwinian selection at a fraction of sites within the gene. The protocols for Activities 3 and 4 are designed to illustrate the application of these LRTs to real gene sequences.

The PAML program

PAML (Phylogenetic Analysis by Maximum Likelihood) is a suite of separate computer programs for analysis of DNA and amino acid sequence alignments (Yang, 2007). The package provides a rich set of sophisticated models for analysis under the maximum likelihood framework. The purpose of these models is to estimate parameters and test hypotheses about an evolutionary process that operated over the branches of a phylogenetic tree. In this framework, the phylogenetic tree is treated as a fixed quantity; hence, if it is not known beforehand, it must be estimated from the data by searching the tree space. A limitation of PAML is that it provides inefficient methods for searching phylogenetic tree space; users are encouraged to utilize programs such as PAUP (Swofford, 2003), GARLI (Zwickl, 2006), or RAxML (Stamatakis, 2006) for more efficient tree searches under the likelihood criterion. The merit of the PAML package is its capacity to analyze the evolutionary processes that give rise to a set of sequences.

The PAML package, including documentation and all source codes, is freely available from *http://abacus.gene.ucl.ac.uk/software/paml.html*. Pre-compiled executable programs are provided for Windows users. Mac users, and users of Unix or Unix-like systems, must compile the programs; detailed instructions are provided online at the above URL. An alternative package for likelihood-based analyses of molecular sequence data is HyPhy (available at *http://www.hyphy.org*). This is an excellent package, and readers are referred to a review by Kosakovsky Pond and Muse (2005) for a good introduction.

The tutorials that follow are based on the `CODEML` program in PAML. A typical analysis will require three plain-text files: (1) a multi-sequence alignment, (2) a tree file, and (3) a control file. The sequence alignment must adhere to certain format conventions referred to as the "modified PHYLIP" format. An example is shown in Figure 19.1.3A, and full details are provided in the PAML documentation. The tree file is a plain-text representation of a phylogenetic tree using parenthetical notation (Fig. 19.1.3B). The `CODEML` program is controlled by variables contained in a control file called `codeml.ctl`. Options that do not apply to a particular analysis can be deleted, and the control files that are provided online for each activity in this unit have been simplified in this way. Detailed descriptions of all the variables for the `CODEML` program are provided in the PAML documentation.



```
5  60

Mouse    ATG GCA ACC CTC AAG GAC CAG CTG ATT GTG AAT CTT CTT AAG GAA GAG CAG GCT --- CCC
Rabbit   ATG GCA GCT CTC AAG GAT CAG CTG ATT CAC AAC CTT CTG AAG GAA GAA CAT GTC --- CCC
Human    ATG GCA ACT CTA AAG GAT CAG CTG ATT TAT AAT CTT CTA AAG GAA GAA CAG ACC --- CCC
Chicken  ATG TCT --- CTC AAG GAT CAT CTC ATC CAC AAT GTC CAC AAA GAG GAG CAC GCT CAT GCC
Lizard   ATG TCT --- CTC AAG GAG AAG CTG ATC GTA AAT GTT CAC AAA GAA GAG CAG CCC CAT GCC
```

```
(((mouse, rabbit), human), (chicken, lizard))
```

plain-text representation of phylogenetic relationships
tree diagram of phylogenetic relationships

Figure 19.1.3 Example of the plain-text representation of (**A**) a multiple sequence alignment in `PAML` format, and (**B**) the phylogenetic relationships for five lineages of vertebrates. A tree diagram is provided for clarity but should not be included in the plain-text file. The alignment comprises the first 20 codons (60 nucleotides) of the *Ldh-A* (lactate dehydrogenase A).

CODON MODELING ACTIVITIES USING THE CODEMI PROGRAM

Activity 1: Maximum likelihood estimation of the intensity of natural selection pressure

The objective of this activity is to use CODEML to evaluate the likelihood of the gamma globin genes from a chimpanzee (*Pan troglodytes*) and a gibbon (*Hylobates lar*) for a variety of ω values. By plotting the log-likelihood scores for each value of ω you will be able to visualize the likelihood surface, and thereby be able to determine the value of ω that maximizes the likelihood of observing the sequence data. This value is the MLE of ω for these data. This MLE is a good estimate of the intensity of natural selection pressure that has acted on these gamma globin gene sequences. You will also check your finding by running CODEML's likelihood optimization algorithm, which you must use to obtain MLEs when you have more sequence data or more complex models.

Protocol

1. Find the online supplementary files (see *http://currentprotocols.com/protocol/mb1901*) for Activity 1 (`A1_codeml.ctl, A1_seqfile.txt`) and familiarize yourself with them. Because the MSA (`A1_seqfile.txt`) contains only a pair of sequences, this activity does not require a tree file. Pay close attention to the modified control file, as it must be edited several times. When ready to run CODEML, delete the "`A1_`" prefix (for this activity the control file must be called `codeml.ctl`).

2. Create a directory where you want your results to go, and place both files (`A1_codeml.ctl, A1_seqfile.txt`) within it. Now open a terminal (or DOS window), move to the directory that contains those files, and run CODEML. Depending on how your system is set up, you might need to place an executable copy of the CODEML program in that same directory. Assuming that CODEML is in the directory that you are working in, run the program by typing "`./codeml`" at the command-line prompt.

3. Familiarize yourself with the results (a file called `A1_HelpFile.pdf` is provided online to assist with this task). If you have not edited `A1_codeml.ctl,` the results will be written to a file called `results.txt`. Identify the line within the results file that gives the likelihood score for the example dataset.

4. Now use a text editor to change the settings of the control file and re-run CODEML. The objective is to compute the likelihood of the example dataset given a fixed value of ω.
   a. Change the name of your result file (via `outfile=` in the control file) or you will overwrite your previous results!
   b. Change the fixed value for ω by changing the value for `omega=` in the control file. The values for this exercise are provided as comments at the bottom of the `A1_codeml.ctl` file.

5. Repeat step 4 for each of the alternate values of ω given at the bottom of `A1_codeml.ctl`.

6. Use a spreadsheet application, or a statistical package, to plot the log-likelihood score (*y*-axis) against the fixed value for ω (*x*-axis). Use a logarithmic scale for the *x*-axis (do not transform ω).

7. Use your plot to try to guess the value of ω that will maximize the likelihood score (i.e., the MLE of ω for the gamma globin sequences).

8. Now change the control file so that CODEML will use its likelihood optimization algorithm to find the MLE by setting `fix_omega=0` in the control file. This means that the value you provide for "`omega =`" will then be used as the starting value for the optimization algorithm. Use `omega=2.00` as your starting value and compare the result to your guess from step 7.

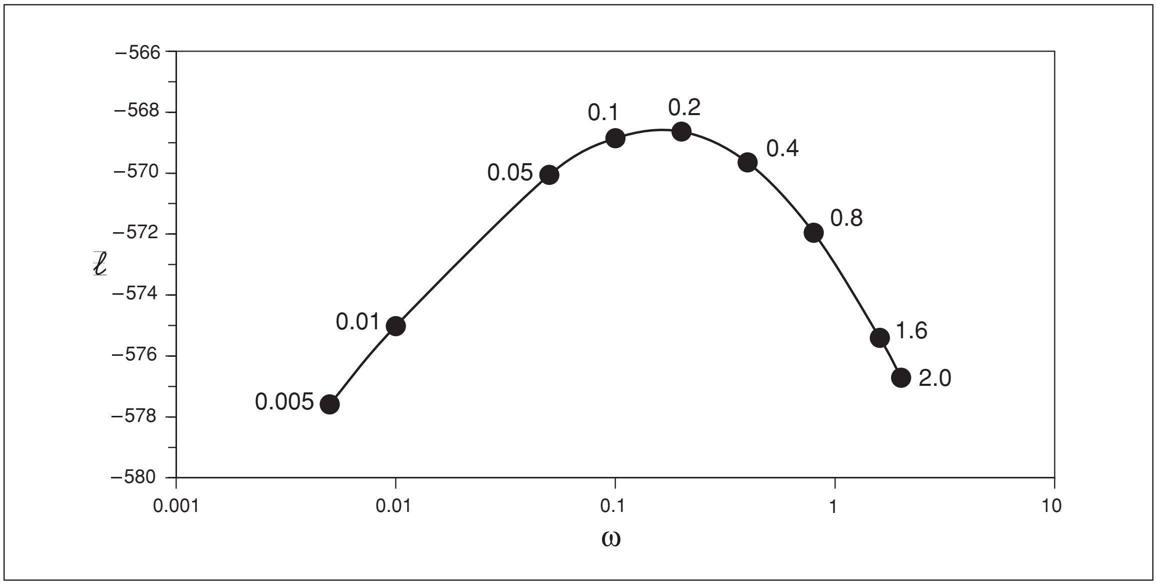


Figure 19.1.4 The log likelihood (ℓ) for the gamma globin genes from a chimpanzee (*Pan troglodytes*) and a gibbon (*Hylobates lar*) as a function of the ω parameter of codon model M0. The maximum likelihood estimate of ω is the value that maximizes the likelihood function (–568.58); for these data it is 0.1623.

Commentary

The plot of likelihood scores obtained in step 6 above should look very similar to the plot provided in Figure 19.1.4. This plot shows the log likelihood as a function of just the ω parameter, but the model contained parameters for the transition/transversion rate ratio (κ) and genetic distance between the chimpanzee and gibbon gene sequences (*t*). The `A1_codeml.ctl` was configured so that MLEs for κ and *t* were re-estimated for every fixed value of ω via an automatic optimization algorithm in `CODEML`. Any plot that describes the likelihood of a single parameter when nuisance parameters (e.g., κ and *t*) are optimized is called a *profile likelihood*.

If the parameters κ and *t* had been fixed rather than optimized, then the process of determining the MLE for ω would have been a simple *univariate optimization*. However, users of codon models will be faced with more complex *multivariate optimization* problems. Codon models are typically applied to datasets comprising many sequences, which are related by a phylogenetic tree. All the branches of that tree will have a length parameter that should be optimized in addition to the parameters of the codon model. Computer programs must be used to solve such optimization problems. Users of the `CODEML` program can choose from two different algorithms. The first (`method=0`) is a full multidimensional optimization that updates all parameters, including branch lengths, at the same time. The second (`method=1`) cycles between a series of fast univariate optimizations of individual branch length parameters and a simultaneous optimization of codon model parameters. Each approach has its pros and cons. Because parameter optimization under a complex model can be very challenging and is not guaranteed to succeed, users should be familiar with the alternatives implemented within the `CODEML` program, and in other programs as well [see Yang (2006) for additional details].

Activity 2: Investigating the impact of assumptions

A codon model, being an explicit representation of a complex evolutionary process, will contain biologically relevant parameters in addition to ω. The way that these parameters are treated can be thought of as the assumptions of the model. Let us take the rate of transitions substitution as an example. If transitions had evolved at the same rate as transversions in a dataset, then it would be sensible to set κ = 1 in the model. The benefit of this treatment of κ would be one less parameter to optimize. However if the assumption of $\kappa = 1$ were incorrect, then the estimates of the other parameters (e.g., $\hat{\omega}$) in the model could be affected (Yang and Nielsen, 2000).

In codon models, the probability of a change from one codon to another is proportional to the frequency of the target codon, which is denoted as π_j in the codon model (Fig. 19.1.1B). Because the codon frequency parameter is a fundamental aspect of the model, the role it plays within the model has received much attention (e.g., Aris-Brosou and Bielawski, 2006; Rodrigue et al., 2008; Yap et al., 2010); some of the more widespread treatments of the π's are reviewed here. The simplest approach is to assume codon usage is unbiased and set all 61 frequency parameters (under the universal genetic code) equal to 1/61 (referred to as "Fequal"). Some genes, such as the *amy2* gene of *Drosophila melanogaster* and *Drosophila pseudoobscura*, can have highly biased codon usage (see Supplementary_FigureS1.pptx; *http://currentprotocols.com/protocol/mb1901*), so the Fequal assumption would seem to be inappropriate for *amy2*. An alternative approach is to treat the frequency of each codon (π_j) as an independent model parameter (referred to as "F61" or "Ftable"). The drawback of this approach is that it adds 60 more parameters to the model. A compromise is to assume that the frequency of each codon could be different, but to estimate each π_j as a simple product of the frequency of nucleotides; for example, $\pi_{\mathrm{ATG}} = \pi_{\mathrm{A}} \times \pi_{\mathrm{T}} \times \pi_{\mathrm{G}}$. This approach (referred to as "F3×4") requires adding just 9 parameters to the model.

The objective of this activity is to use CODEML to investigate the sensitivity of $\hat{\omega}$ to assumptions about κ and π's. CODEML provides estimates of several different measures of sequence evolution, and these can be used to explore the impact of model assumptions. Users should pay close attention to the estimates of number of synonymous (*S*) and nonsynonymous sites (*N*), as well as the rates of synonymous (d_{S}) and nonsynonymous (d_{N}) substitution. In this activity, the impact of the model assumptions can be explored with respect to *S*, d_{S}, $\hat{\omega}$, and ℓ. The dataset will be an alignment of 493 codons from the *amy2* gene of *D. melanogaster* and *D. pseudoobscura*.

Protocol

1. Find the online supplementary files (*http://currentprotocols.com/protocol/mb1901*) for Activity 2 (`A2_codeml.ctl, A2_seqfile.txt`) and familiarize yourself with them. It might be convenient to create a new directory for Activity 2. When ready to run CODEML, delete the "`A2_`" prefix from the filenames.

2. Run CODEML using the settings in the control file provided for Activity 2. Familiarize yourself with the results (the file `A2_HelpFile.pdf` will help with this). In addition to identifying the log-likelihood score (ℓ), κ, and $\hat{\omega}$, you must be able to identify the part of the result file that provides estimates of the following:
   a. Number of synonymous or nonsynonymous sites (*S* and *N*).
   b. Synonymous and nonsynonymous rates (d_{S} and d_{N}).

3. As in Activity 1, change the control files and re-run CODEML. The objective is to compute the likelihood of the example dataset under different model assumptions. To do this it is necessary to perform the following steps:
   a. Change the name of the main result file (via `outfile=` in the control file) or else the previous results will be overwritten.
   b. Change the model assumptions about codon frequencies (via `CodonFreq=`) and kappa (via `kappa=` and `fix_kappa=`).
   c. Repeat steps 3a and 3b for each set of assumptions about codon frequencies and kappa given below (and as comments at the bottom of the example control file).

i. `CodonFreq=0; kappa=1; fix_kappa=1`

ii. `CodonFreq=0; kappa=1; fix_kappa=0`

iii. `CodonFreq=2; kappa=1; fix_kappa=1`

iv. `CodonFreq=2; kappa=1; fix_kappa=0`

v. `CodonFreq=3; kappa=1; fix_kappa=1`

vi. `CodonFreq=3; kappa=1; fix_kappa=0`

4. Create a table to store results for κ, S, N, d_S, d_N, $\hat{\omega}$, and ℓ.

5. Determine which model assumptions yield the largest and smallest values of S, d_S, and $\hat{\omega}$, and which assumptions maximize the likelihood of observing the *amy2* gene sequences.

Commentary

The statistics estimated from the *amy2* sequences in Activity 2 should be very close to those presented in Table 19.1.1. Results indicate that model-based inferences will be sensitive to assumptions. Ignoring transition bias (assuming $\kappa = 1$) has a minor effect. Estimates of the number of synonymous sites (S) are sensitive to this assumption, but the effect is small compared to the impact of assumptions about the π's. Estimates of S are profoundly affected when codon frequencies are incorrectly assumed to be unbiased (Fequal), with estimates being substantially larger than when codon bias is included in the model. The consequent effect on d_S is important, because under Fequal the values of d_S are not consistent with saturation of substitutions ($d_S < 1$), whereas under the more complex F61 model the estimates of d_S indicate that synonymous changes have become saturated ($d_S > 2$). Assumptions about the π's also impact the conclusions about selection intensity. Estimates of $\hat{\omega}$ differ as much as 6-fold, with negative selection pressure appearing moderately strong ($\hat{\omega}$~0.17) under Fequal and very strong ($\hat{\omega}$~0.02) under F61. Given that codon frequencies are very biased (see Supplementary_FigureS1.pptx; *http://currentprotocols.com/protocol/mb1901*) and that F61 π's provide a significantly better fit to these data (LRT not shown), estimates under F61 are expected to be more reliable.

Activity 2 explores just the more common treatments of the π's within codon models. Other treatments can be warranted. For example, if biases in codon frequencies arise

Table 19.1.1 Sensitivity of Results for the *amy2* Sequences to Different Assumptions Within the Codon Model[a]

Model	κ[b]	S[c]	d_S[d]	N[e]	d_N[f]	$\hat{\omega}$[g]	ℓ[h]
Fequal, $\kappa = 1$	1	376.8	0.3671	1102.2	0.0644	0.1754	−2693.29
Fequal, κ estimated	0.94	373.7	0.3709	1105.3	0.0642	0.1731	−2693.22
F3×4, $\kappa = 1$	1	147.0	1.2467	1332.0	0.0559	0.0448	−2430.98
F3×4, κ estimated	1.33	148.4	1.2182	1330.6	0.0561	0.0460	−2429.90
F61, $\kappa = 1$	1	130.1	2.2452	1348.9	0.0553	0.0246	−2307.16
F61, κ estimated	1.25	132.5	2.0639	1346.5	0.0555	0.0269	−2306.72

[a] The statistics estimated in Activity 2 should be very close to those presented in this table.
[b] κ is the transition/transversion rate ratio.
[c] S is the number of synonymous sites.
[d] d_S is the synonymous substitution rate.
[e] N is the number of nonsynonymous sites.
[f] d_N is the nonsynonymous substitution rate.
[g] $\hat{\omega}$ is the maximum likelihood estimate of the d_N/d_S ratio.
[h] ℓ is the log-likelihood score.

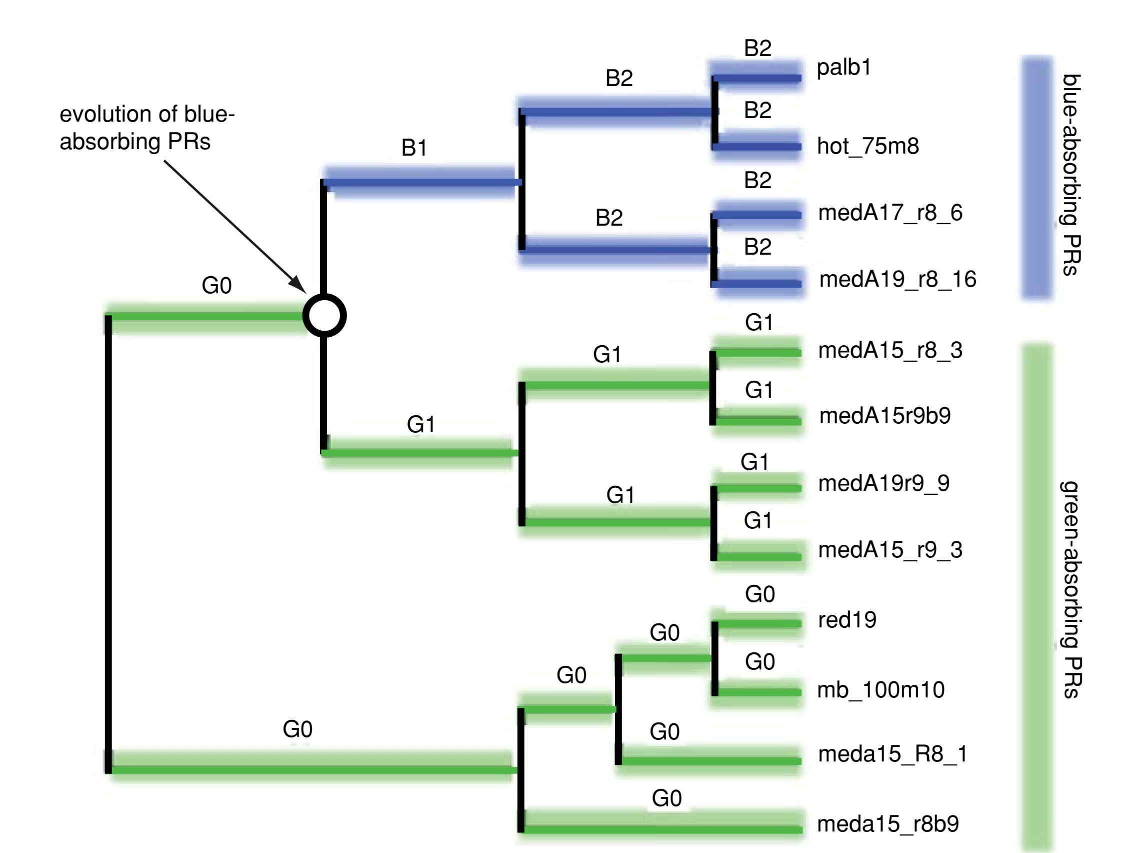
predominantly from the mutational process acting at the level of the DNA sequence, then codon frequencies might be better modeled as proportional to the equilibrium frequency of the target nucleotide (Muse and Gaut, 1994) rather than that of the target codon. Pathogen genomes sometimes exhibit a more extreme form of codon bias, and the transition probability might be best modeled by the frequency of the target nucleotide conditioned on the nucleotide states at the other two sites in the codon (Yap et al., 2010). Codon usage can sometimes vary substantially among sites within a single gene, such as among those sites that encode the helices and loops of transmembrane proteins, requiring a codon model that permits the π's to vary among sites (Bao et al., 2007). Failure to appropriately model the more important nuisance parameters, such as the π's, will lead to biases in $\hat{\omega}$ and possibly false biological conclusions (Aris-Brosou and Bielawski, 2006; Bao et al., 2008; Yap et al., 2010). Activity 2 illustrates the importance of exploratory analysis of model assumptions.

Activity 3: Testing hypotheses about variation of selection pressure over evolutionary time

A pairwise analysis, as in Activities 1 and 2, must average the estimate of ω over the entire evolutionary history that separates the pair of sequences. However, the intensity of natural selection can change over evolutionary time (Anisimova and Liberles, 2012). To detect such a pattern of evolution, codon models must be applied to a well-sampled phylogenetic tree so that different ω parameters can be specified for different branches of the tree (Yang, 1998). Such an approach, often referred to as a "branch model" for ω, does increase model complexity. The benefit is that users can specify nested sets of models for the purpose of testing evolutionary hypotheses via the LRT. Generally useful examples include the LRT for altered selective intensity along a single branch (i.e., testing the fit of an episodic model of evolution), or the LRT for altered selection pressure within a specific clade (i.e., testing the fit of a long-term shift model). The episodic model and the shift model have been valuable in studies of biological events that have occurred at particular points in evolutionary history such as following a gene duplication event (e.g., Bielawski and Yang, 2001), following the transmission of parasitic organism to a new host (e.g., Jiggins et al., 2002), or during the transition to a new ecological niche by a free-living species (e.g., Kelley and Swanson, 2008).

The objective of this activity is to employ branch models to explicitly test hypotheses about the evolutionary history of proteorhodopsin (PR) gene sequences. The PR protein is a retinal-binding bacterial protein that functions as a light-sensitive proton pump (DeLong and Béjà, 2010). It is widespread in marine bacteria, and it represents a globally significant source of light-based metabolism in the world's open oceans (DeLong and Béjà, 2010). Genetic divergence among PRs is, in part, associated with fine-tuning the spectral sensitivity of the protein to match different light intensities in different oceanic habitats (e.g., depth). This activity will focus on the evolutionary divergence of blue-absorbing PRs from a green-absorbing ancestor, which can easily be traced to a particular node of the PR gene tree (Fig. 19.1.5). Three hypotheses about the history of selection pressure (Fig. 19.1.5) will be tested via three LRTs. We begin with the null hypothesis (H_0): the intensity of selection pressure (ω) is the same for all branches of the PR gene tree. We are interested in alternatives to H_0 where selection pressure might not have been constant over time. These must be formulated prior to analyzing the data, and in this case we will investigate one episodic model and two shift models:

H_1: A family-wide shift in selection occurred (both in blue- and green-absorbing PRs) following the evolution of blue-absorbing PRs (Fig. 19.1.5). Since H_0 is a case of H_1 (Fig. 19.1.5), a LRT with 1 degree of freedom (df) can be used to test H_0 against H_1. Acceptance of H_1 by this test is evidence that selection pressure changed sometime after the evolution of blue-absorbing PRs.



$$H_0\text{: } \omega_{G0} = \omega_{G1} = \omega_{B2} = \omega_{B1}$$
$$H_1\text{: } \omega_{G0} \neq \omega_{G1} = \omega_{B2} = \omega_{B1}$$
$$H_2\text{: } \omega_{G0} = \omega_{G1} \neq \omega_{B2} = \omega_{B1}$$
$$H_3\text{: } \omega_{G0} = \omega_{G1} = \omega_{B2} \neq \omega_{B1}$$

Figure 19.1.5 Phylogeny for 12 PR gene sequences. Branch lengths are not to scale. The one-ratio model (H_0) assumes the same intensity of selection pressure over all branches. The H_1 model is based on the hypothesis that a family-wide shift in selection (both the blue- and green-absorbing PRs) occurred following the evolution of blue-absorbing PRs. H_1 assumes that selection intensity prior to this event (ω_{G0}) differs from selection intensity after this event ($\omega_{G1} = \omega_{B1} = \omega_{B2}$). H_2 models a shift in selection pressure only within the blue-absorbing PRs. Hence, selection intensity is uniform for all blue-absorbing PRs ($\omega_{B1} = \omega_{B2}$) and differs from all green-absorbing PRs ($\omega_{G0} = \omega_{G1}$). H_3 is based on an episodic model of functional divergence where altered selection occurred only in the branch associated with the evolution of blue-absorbing PRs (ω_{B1}), with all other branches being subject to the ancestral levels of selection intensity ($\omega_{G0} = \omega_{G1} = \omega_{B2}$). For the color version of this figure, go to *http://currentprotocols.com/protocol/mb1901*.

H_2: A shift in selection occurred only in the blue-absorbing PRs (Fig. 19.1.5). An LRT with 1 df can be used to test H_0 against H_2. Acceptance of H_2 by this test is evidence for unique selection pressure acting on the blue-absorbing PRs, with selection pressure acting on the ancestral green-absorbing PRs being the same as that acting on green-absorbing PRs that evolved after the evolution of the blue-absorbing PRs.

H_3: An episode of altered selection occurred only within the branch that immediately follows the origin of the blue-absorbing PRs (Fig. 19.1.5). A LRT with 1 df can be used to test H_0 against H_3. Acceptance of H_3 by this test is evidence for functional divergence of blue-absorbing PRs by episodic natural selection, with all PRs (blue- and green-absorbing) subject to the ancestral level of selection pressure after the origination event.

Each alternative hypothesis (H_1, H_2, and H_3) is constructed as a branch model in CODEML, with the details of the branch-specific ω parameters being supplied by the user via annotations to a tree file. Tree files pre-annotated for H_1, H_2, and H_3 are provided via the online supplementary materials. Note that the PAML manual provides a detailed description of tree file formatting and annotation. A dataset and an annotated control file are also provided so that users following the protocol can fit the branch models required to carry out all three LRTs.

Protocol

1. Obtain the online supplementary files (*http://currentprotocols.com/protocol/mb1901*) for Activity 3 (`A3_codeml.ctl, A3_seqfile.txt, treeH0.txt, treeH1.txt, treeH2.txt, treeH3.txt`). The tree files represent different hypotheses denoted H_0, H_1, H_2, and H_3. Figure 19.1.5 illustrates how independent ω parameters are specified for different branches of the tree for each hypothesis.

2. When ready to run CODEML, delete the "A3_" prefix from the filenames. Run CODEML using the settings in the control file and familiarize yourself with the results (the file `A3_HelpFile.pdf` will help with this). In addition to identifying the likelihood score, you must be able to identify the branch-specific estimates of the ω parameter. As this is the first run of CODEML, the branch model is H_0 and all values for ω will be the same. In later runs there will be differences among ω values for some branches.

3. Change the control files and re-run CODEML to fit a branch model (H_1, H_2, or H_3) to the data. Because the relevant model information is contained in the tree file, you will need to edit the control file so that it reads the appropriate tree file.

   a. As always, change the name of the main result file (via `outfile=` in the control file) to avoid overwriting your previous results.

   b. Change the model assumptions about branch-specific ω values by changing the tree file (via `treefile=`) and set `model=1` within the control file.

   c. Run CODEML and collect the results for ω and ℓ.

   d. Repeat steps 3a to 3c for each of the tree files `treeH1.txt, treeH2.txt`, and `treeH3.txt`.

4. Create a table to store results for ω and ℓ. Then, carry out the LRTs a to c given below. The test statistic for each LRT is $2 \times (\ell_{\text{alt}} - \ell_0)$.

   a. H_1 vs. H_0 (1 df).

   b. H_2 vs. H_0 (1 df).

   c. H_3 vs. H_0 (1 df).

Use the program CHI2 provided in the PAML package to determine the *P*-value for each test statistic. Note that CHI2 takes two arguments, the df and the test statistic. To invoke this program, type "`./chi2 p`" at the command line and then provide values for the df and the test statistic when prompted.

Commentary

Activity 3 should yield results very similar to those presented in Table 19.1.2. The first LRT is significant (H_1 vs. H_0: $P = 0.015$), suggesting that the average intensity of selection pressure had changed after the evolution of blue-absorbing PRs. Inspection of the parameter estimates under H_1, however, reveals that the average selection pressure after the evolution of blue-absorbing PRs ($\hat{\omega} = 0.047$) is only slightly different than the ancestral level ($\hat{\omega} = 0.076$). Thus, this test provides evidence for only a subtle change in the intensity of purifying selection pressure. If the intensity of selection changed in only a subset of branches following the evolution of blue-absorbing PRs, then the intensity

Table 19.1.2 Parameter Estimates and LRTs for Models of Variable Selection Pressure (ω) Among Lineages of PR Proteins[a]

Model[b]	ω_{G0}	ω_{G1}	ω_{B2}	ω_{B1}	ℓ[c]	LRT[d]
$H_0 : \omega_{G0} = \omega_{G1} = \omega_{B2} = \omega_{B1}$	0.0625	$= \omega_{G1}$	$= \omega_{B2}$	$= \omega_{B1}$	–3622.84	NA
$H_1 : \omega_{G0} \neq \omega_{G1} = \omega_{B2} = \omega_{B1}$	0.0470	$\neq 0.0765$	$= \omega_{B2}$	$= \omega_{B1}$	–3619.88	$P = 0.015$
$H_2 : \omega_{G0} = \omega_{G1} \neq \omega_{B2} = \omega_{B1}$	0.0558	$= \omega_{G1}$	$\neq 0.0730$	$= \omega_{B1}$	–3621.94	$P = 0.180$
$H_3 : \omega_{G0} = \omega_{G1} = \omega_{B2} \neq \omega_{B1}$	0.0570	$= \omega_{G1}$	$= \omega_{B2}$	$\neq 0.1739$	–3619.27	$P = 0.007$

[a] The statistics estimated in Activity 3 should be very close to those presented in this table.

[b] The topology and branch-specific ω parameters for each model are presented in Figure 19.1.5.

[c] ℓ is the log-likelihood score under the model.

[d] The LRTs are as follows: H_0 vs. H_1; H_0 vs. H_2; H_0 vs. H_3. All LRTs have df = 1. NA, not applicable.

measured as an average over all branches, as modeled in H_1, might have been diluted. The additional LRTs are used to investigate such scenarios. The second LRT is based on the hypothesis that a shift in selection intensity might have occurred only within the clade of blue-absorbing PRs. As this LRT is not significant (H_2 vs. H_0: $P = 0.18$), we have no evidence for this model of functional divergence in these data. The third LRT is based on the hypothesis that an episode of adaptive evolution occurred immediately following the evolution of the blue-absorbing PRs. This LRT is significant (H_3 vs. H_0: $P = 0.007$), and parameter estimates under H_3 give the largest difference in selection intensity among the alternative models (Table 19.1.2). At face value, $\hat{\omega}$'s under the H_3 model indicate a relaxation of selection intensity immediately after the evolution of the blue-absorbing PRs ($\hat{\omega}_{B2} = 0.1739$). However, this averages ω over all the sites in the PR gene, and we expect that most of those sites would not have been subject to altered selection pressures and would have been dominated by purifying selection. Thus the elevated estimate in the single branch could reflect the presence of a small fraction of sites evolving under positive Darwinian selection pressure ($\omega > 1$). Further analysis of an expanded dataset, with additional models, revealed this to be the case (Bielawski et al., 2004).

Optimization of the model parameters in Activity 3 will have taken noticeably more time than it did in Activities 1 and 2. This is because Activity 3 is based on a phylogeny having 21 branches, and the length of each branch must be optimized in addition to the parameters for κ and ω in the model. Codon models can be made even more complex, but any added complexity must be carefully considered. For example, it is possible to create a model having an independent ω parameter for every branch in the tree. This so-called "free ratios" model is too complex for the PR dataset as well as most others. Those seeking to discover variation in ω without having to specify the lineages a priori are encouraged to consider the genetic algorithm of Kosakovsky Pond and Frost (2005a) or the clustering algorithm of Dutheil et al. (2012).

Activity 4: Testing for variability of selection pressures among sites

The individual amino acid sites of a protein are subject to a wide spectrum of selective pressures. For genes encoding a functional protein, the majority of codon sites will have been subjected to strong purifying selection. In the rare cases when protein evolution has been adaptive, we expect that only a small subset of the codon sites would have been evolving under positive Darwinian selection ($\omega > 1$). For this reason, models that accommodate variable selection pressures among sites have more power to detect positive selection than models that permit selection pressure to vary among branches. Models that permit selection to vary among sites are often referred to as "sites-models" and have

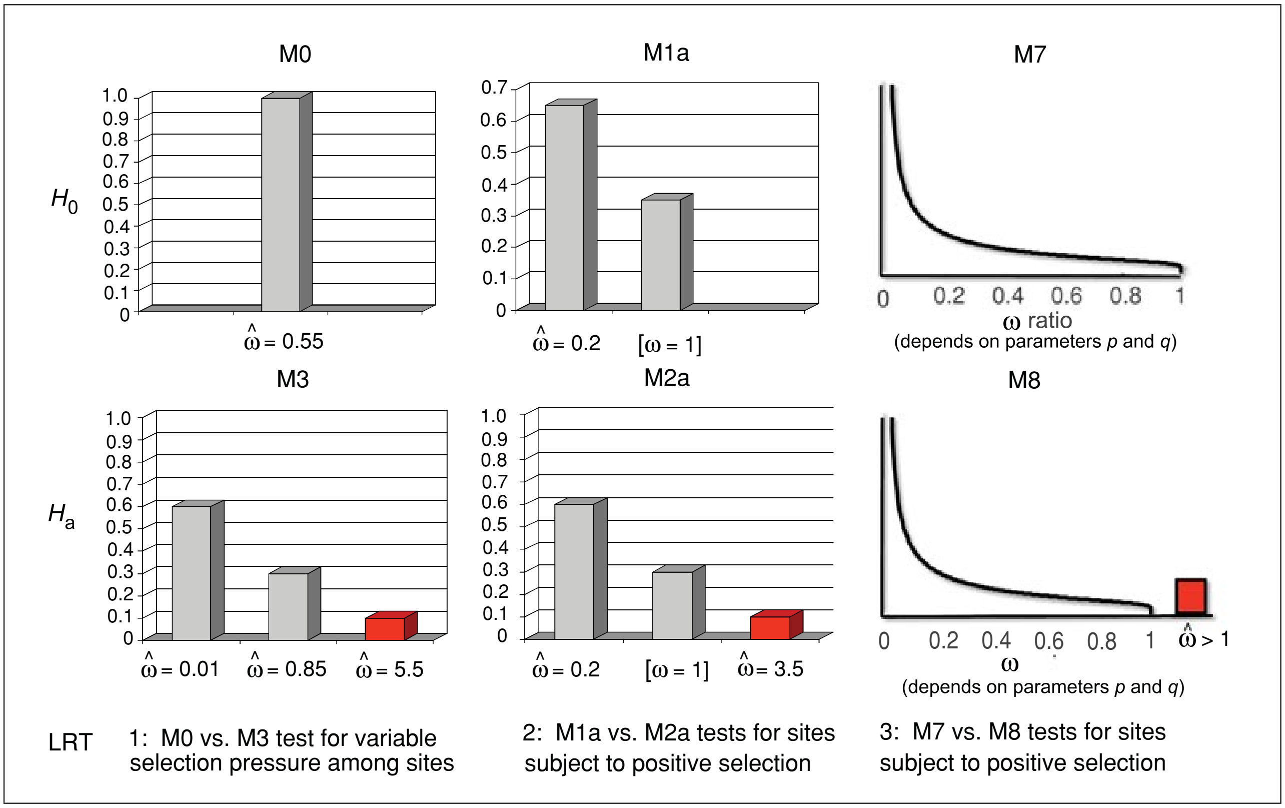


Figure 19.1.6 Graphical representation of the ω distributions of codon models M0, M1a, M2a, M7, and M8. The nested relationships between M0 vs. M3, M1a vs. M2a, and M7 vs. M8 permit three different likelihood ratio tests (LRTs). The LRT of M0 vs. M3 is a test for variable selection pressure among sites and has df = 4 when M3 has three discrete categories for ω. The LRTs of M1a vs. M2a and of M7 vs. M8 are alternative tests for a fraction of sites subject to positive selection. Each of these LRTs has df = 2. For the color version of this figure, go to *http://currentprotocols.com/protocol/mb1901*.

been the subject of considerable development and testing (e.g., Anisimova et al., 2001; Kosakovsky Pond and Frost, 2005b; Bao et al., 2008).

The objective of Activity 4 is to use a recommended set of six sites-models (Fig. 19.1.6) to test for variability in selection pressure among sites. These models employ a statistical distribution to permit ω to vary among sites, and can be used to carry out three different LRTs. Other approaches are possible, but will not be employed here. The models are M0, M1a, M2a, M3, M7, and M8. This set of models is recommended because the involved LRTs (Fig. 19.1.6) are well studied and have generally desirable statistical properties (Anisimova et al., 2001; Bao et al., 2008). Pairs of these models are used to carry out three LRTs:

LRT-1: Models M0 and M3 are used to test for variable selection intensity among sites. Model M0 is the null model, and it assumes the same selection pressure (a single ω) for all sites. Model M3 assumes that sites fall into k discrete categories of evolution, each having a parameter for selection (ω_i) and the proportion of sites (p_i) in the gene. In this activity, k is set to 3. This LRT has 4 degrees of freedom.

LRT-2: Models M1a and M2a comprise a test for sites subject to positive selection. M1a is the null model of this LRT, and it specifies just two classes of sites: conserved ($\omega < 1$) and neutral ($\omega = 1$). M2a forms the alternative model by adding a third class of sites dedicated to positive selection ($\omega > 1$). This LRT has 2 degrees of freedom.

LRT-3: Models M7 and M8 comprise an alternative test for sites subject to positive selection. Model M7 uses a flexible β distribution to permit ω to vary among sites.

M7 is the null model because the β distribution is restricted to the interval (0,1). For computational convenience, the β distribution is divided into 10 discrete bins. M8 adds an extra discrete category to M7 which has ω that is free to take a value >1. This LRT also has 2 degrees of freedom.

All models except M0 assume that ω varies among sites, but they do not require knowledge of which sites belong to the different model categories for ω. Collectively, these are referred to as "mixture models." Further details about the mixtures for ω in these models are provided in Yang et al. (2000). Fitting these models involves estimating both ω_i and their mixing parameters (p_i) by maximum likelihood. Note that the LRT based on M1a and M2a is more conservative than the LRT based on M7 and M8, but this comes with a cost, as the former LRT is less powerful. A gene is considered to have a signature of positive selection if (1) an appropriate LRT is significant and (2) at least one of the MLEs of ω is greater than 1. Given such evidence, an empirical Bayesian technique can be used to infer which sites had been subject to positive selection.

The dataset for Activity 4 is a set of 15 gene sequences from the bacterium *Flavobacterium psychrophilum*. These sequences encode a family of cell-surface proteins that were identified during a large-scale survey of genome evolution among the Bacteroidetes. The gene family is characterized by a high rate of evolution, and *F. psychrophilum* is a known pathogen. Thus, a history of evolution by diversifying positive selection is plausible for these sequences. This activity demonstrates how sites-models are employed to statistically test for positive selection, and to infer the individual sites that were the target of such evolution.

Protocol

1. Obtain the online supplementary files (*http://currentprotocols.com/protocol/mb1901*) for Activity 4 (`A4_codeml.ctl`, `A4_seqfile.txt`, `treeM0.txt`, `treeM1.txt`, `treeM2.txt`, `treeM3.txt`, `treeM7.txt`, `treeM8.txt`). Each tree file is pre-loaded with the MLEs for the branch lengths. Pay close attention to the modified control file, as it must be edited to specify models M1a, M2a, M3, M7, and M8.

2. If you plan to run two or more models at the same time, then create a separate directory for each run and place a sequence file, control file, and tree file in each directory.

3. As in all the previous activities, delete the "`A4_`" prefix, edit the control file, and run `CODEML`. The objective is to fit all six codon models (M0, M1a, M2a, M3, M7, and M8) to the example dataset.

   a. When running analyses sequentially in the same directory, change the name of the main result file (via `outfile=` in the control file), or else previous results will be overwritten.

   b. Set the tree file with `treefile=`. Pre-loading the tree files with the MLEs of the branch lengths for each model greatly shortens the run time for this activity, but it requires that a different tree file must be set for each model. See the example control file for additional notes about tree file names for this activity.

   c. Set the codon model with `NSsites=`.

   d. For some models you will also need to set the number of categories (`ncatG`) in the ω distribution:

      i. For M3, set `ncatG=3`

      ii. For M7, set `ncatG=10`

      iii. For M8, set `ncatG=10`

e. In addition to the primary outfile, `CODEML` will produce a supplementary result file called "`rst`". Once the analysis is complete, rename the `rst` file because subsequent runs will overwrite it!

f. Repeat steps 3a through 3e for each of the six codon models.

4. Create a table to store results for the MLEs and the likelihood scores obtained for each model. The file `A4_HelpFile.pdf` will help you to find the MLEs for the ω distribution used by each model.

5. In addition, carry out the following LRTs:

   a. M0 vs. M3 (4 df).

   b. M1a vs. M2a (2 df).

   c. M7 vs. M8 (2 df).

6. Lastly, open the `rst` file generated when you ran model M3 and review the output. The file `A4_rst_HelpFile.pdf` will help you to identify the posterior probability that each site evolved under positive selection ($\omega > 1$). Although output to the `rst` file varies somewhat according to the model, the last two columns of the naïve empirical Bayes (NEB) and Bayes empirical Bayes (BEB) results are of general interest. The last column gives the posterior probability that a site evolved under $\omega > 1$. Large values, such as $\geq$0.95, are desirable. The second-to-last column gives the posterior mean value for ω on a per-site basis. This site-specific estimate of ω depends on the posterior probabilities provided for that site in the preceding columns of the output. When searching for positively selected sites, a posterior mean $\omega > 1$ is noteworthy.

7. Empirical Bayes is used to classify sites according to ω. Use the help file to find and plot the posterior probability under each category for ω in the model. Figure 19.1.7 provides an example of how such plots are used to profile the distribution of selection pressures acting on amino acid sites. Consult the online supplementary materials (*http://currentprotocols.com/protocol/mb1901*) for a summary of the empirical Bayes method.

Commentary

Parameter estimates and likelihood scores for this family of cell-surface proteins are given in Table 19.1.3. Note that averaging ω over sites, as done by M0, indicates that purifying selection dominates this gene ($\hat{\omega} = 0.30$). However, this gene is subject to variable selection pressure among sites, as the LRT of M0 against M3 is highly significant ($2\Delta\ell = 301.3$, $P < 0.0001$). MLEs under the models that allow for positive selection pressure (M2a, M3, and M8) indicate that a fraction of such sites are indeed present in these data (Table 19.1.3). Two LRTs can be employed to formally test this hypothesis. Since both LRTs are highly significant (M1a vs. M2a: $2\Delta\ell = 34.96$, $P < 0.0001$; M7 vs. M8: $2\Delta\ell = 34.02$, $P < 0.0001$), the conclusion of positive selection appears robust to the details of the models. MLEs of the proportion of sites under positive selection ($p_{\omega>1}$) suggest that about 4% to 5% of such sites exist within this gene family. These results illustrate how models that average ω over sites can fail to detect genes that have evolved by positive selection.

While the LRTs and MLEs of Activity 4 have generated evidence for positive selection, these results do not inform us about which sites have $\omega > 1$. A third approach, empirical Bayes, is employed to infer this. The Bayes rule is used to compute the posterior probability that each site had evolved, in turn, under each category for ω in the model (explained in the online supplementary materials; *http://currentprotocols.com/protocol/mb1901*). Posterior probabilities for all site classes under M3 are shown for each codon in Figure 19.1.7. Figure 19.1.7 reveals a rapidly evolving region (codons 30 to 125) and a conserved region (codons 126 to 190), although there is variability in

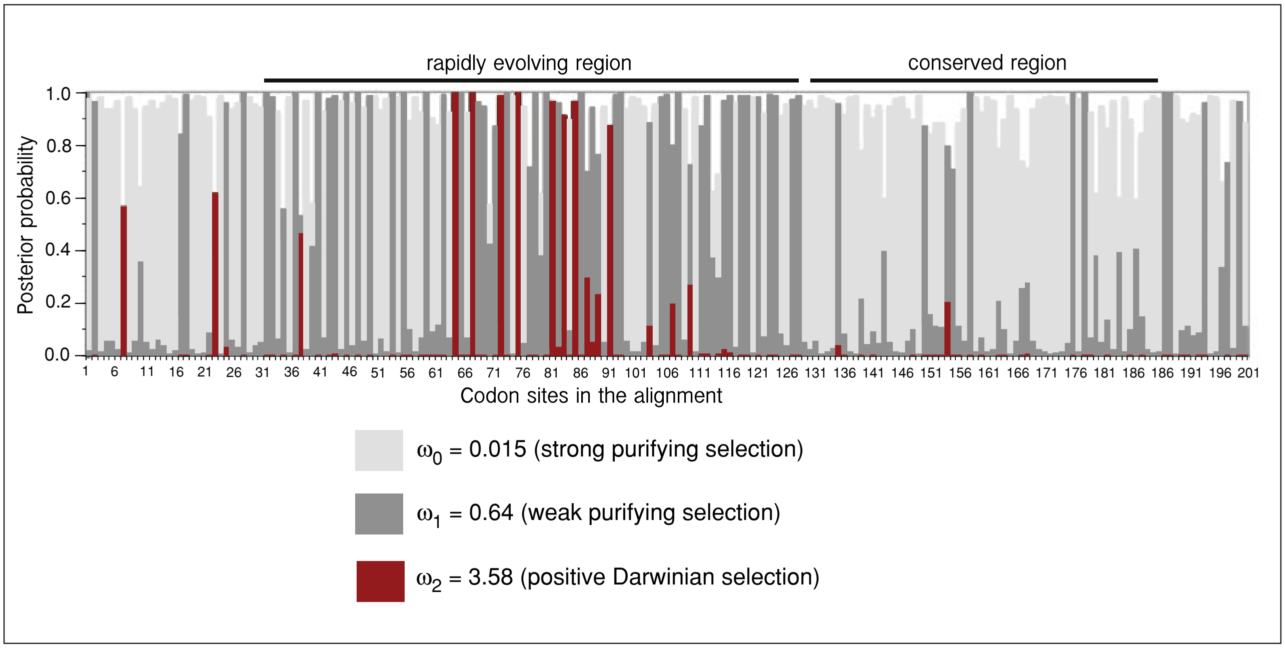


Figure 19.1.7 Posterior probability of ω_0, ω_1, and ω_2 for every site in an alignment of cell-surface proteins from *Flavobacterium psychrophilum*. The posterior probabilities are conditioned on the MLEs for ω under codon model M3. At each site the posterior probabilities sum to 1. Site 1 is an example of codons inferred to have evolved under strong purifying selection; this site has a high posterior probability of $\omega_0 = 0.015$ (posterior probability = 0.98), and a low posterior probability of $\omega_1 = 0.64$ (0.02) and $\omega_2 = 3.58$ (0.0). Site 65 is an example of codons having evolved under positive selection; this site has a high posterior probability of $\omega_2 = 3.58$ (0.998). Note that the data at some sites provide a less conclusive signal (e.g., site 38). For the color version of this figure, go to *http://currentprotocols.com/protocol/mb1901*.

Table 19.1.3 Parameter Estimates and Likelihood Scores for a Set of Cell-Surface Proteins from *Flavobacterium psychrophilum* Under Different Codon Models[a]

Model	NP[b]	ω[c]	ω distribution parameter estimates[d]	PSS[e]	ℓ[f]
M0 (one ratio)	1	0.30	$\omega = 0.30$	None	–3097.56
M3 $k = 3$ (discrete)	5	0.43	$\omega_0 = 0.015$; $\omega_1 = 0.64$; $\omega_2 = 3.58$; $p_0 = 0.60$; $p_1 = 0.35$; $[p_2 = 0.05]$	10 (5)	–2946.91
M1a (nearly neutral)	2	0.37	$\omega_0 = 0.029$; $[\omega_1 = 1.0]$; $p_0 = 0.65$; $[p_1 = 0.35]$	Not allowed	–2969.62
M2a (positive selection)	4	0.54	$\omega_0 = 0.029$; $[\omega_1 = 1.0]$; $\omega_2 = 5.32$; $p_0 = 0.63$; $p_1 = 0.33$; $[p_2 = 0.04]$	7 (3)	–2952.14
M7 (beta)	2	0.30	$p = 0.129$; $q = 0.304$	Not allowed	–2964.37
M8 (beta + ω)	4	0.44	$p = 0.143$; $q = 0.372$; $\omega = 4.24$; $p_0 = 0.96$; $[p_1 = 0.04]$	8 (4)	–2947.36

[a] The statistics estimated in Activity 4 should be very close to those presented in this table.
[b] NP denotes the number of parameters in the model.
[c] The column of single values of ω gives the average over all sites in the PR alignment.
[d] Parameters in the ω distribution within brackets are not free parameters and are presented for completeness.
[e] PSS denotes positively selected sites at a 50% (95%) posterior probability cutoff.
[f] ℓ is the log-likelihood score under the model.

the intensity of selection (and thus evolutionary rate) within these regions. In particular, Figure 19.1.7 indicates that a cluster of 8 sites with high posterior probabilities (>0.90) of $\omega > 1$ are located within the rapidly evolving region (shown in red). Sites discovered in this way are often the subject of further investigation, such as mapping to the 3D structure of the protein, or site-directed mutagenesis and biophysical analysis (e.g., Field et al., 2006).

CONCLUDING REMARKS

The activities and protocols provided here are intended to provide a foundation to build upon. The activities illustrate the core tasks of (1) parameter estimation, (2) hypothesis testing, and (3) site classification. These tasks have different levels of difficulty. Hypothesis testing via the LRT is the easiest. Reliable inferences can even be made from small datasets via LRTs (Anisimova et al., 2001). The LRT based on M1a vs. M2a is noteworthy because it tends to be conservative without losing too much power (Anisimova et al., 2001) and appears quite robust to inappropriate modeling of nuisance parameters (Bao et al., 2008). The next task, parameter estimation, is more difficult. MLEs will often have large amounts of uncertainty, and they can be strongly impacted by inappropriate treatment of other parameters. Parameters of the ω distributions employed in the mixture models are well known for their difficulty to estimate precisely. The third, and most difficult, task is Bayesian identification of sites subject to positive selection. The involved posterior probabilities are sensitive to the details of the model, the quality of the MLEs, and the amount of information that can be contained by the data at just one site. Despite this, prediction of positively selected sites can be reliable and powerful when applied to alignments that contain large numbers of sequences (Anisimova et al., 2002; Kosakovsky Pond and Frost, 2005b).

Users of codon models are encouraged to carry out robustness analyses, keeping in mind the relative difficulty of the involved analytical tasks. Robustness of LRT results to alternative formulations of the models should be investigated wherever possible. If inferences at this level are sensitive to how the models are formulated, then parameter estimates and site classifications are likely to be sensitive as well. Because optimization is not guaranteed to succeed, analyses should be run multiple times and checked for consistency. Any results derived from suboptimal solutions should be discarded. Sensitivity to tree topology also should be explored. Tree topologies are typically inferred from the data in hand and have their own sampling errors. Alternative methods can be used to estimate trees from a single dataset, and the impact of differences among these trees should be explored. As illustrated in Activity 2, users also can explore the robustness of their results to how the nuisance parameters, such as codon bias, are treated within a codon model. Users of programs like CODEML must be actively engaged with the process of modeling the data in hand; what works well for one dataset could be inappropriate for another.

Activities 3 and 4 illustrate how averaging ω over branches or over sites can lead the user of a codon model to miss important features of gene sequence evolution. It follows that a short episode of positive selection acting at just a small fraction of sites could be missed by both the branch models (Activity 3) and the sites-models (Activity 4). A more complex class of codon models, called branch-site models, has been developed to address such cases (Bielawski and Yang, 2004; Zhang et al., 2005; Kosakovsky Pond et al., 2011). Users are encouraged to gain some experience with branch models and sites-models before trying the more complex models. In the appropriate setting, and with sufficiently large samples of data, the branch-site models have proven useful (Yang and dos Reis, 2011). Because such state-of-the-art models push the boundaries of analysis, they are beyond the scope of the protocols presented here. However, the issues surrounding the three core tasks, as well as the notion of robustness, are just as relevant.

LITERATURE CITED

Anisimova, M. and Kosiol, C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.* 26:255-271.

Anisimova, M. and Liberles, D. 2012. Detecting and understanding natural selection. *In* Codon Evolution: Mechanisms and Models (G. Cannarozzi and A. Schneider, eds.) Oxford University Press, New York.

Anisimova, M., Bielawski, J.P., and Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18:1585-1592.

Anisimova, M., Bielawski, J.P., and Yang, Z. 2002. Accuracy and power of Bayesian prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* 19:950-958.

Aris-Brosou, S. and Bielawski, J.P. 2006. Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. *Gene* 378:58-64.

Bao, L., Gu, H., Dunn, K.A. and Bielawski, J.P. 2007. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. *BMC Evol. Biol.* 7:S5.

Bao, L., Gu, H., Dunn, K.A., and Bielawski, J.P. 2008. Likelihood Based Clustering (LiBaC) for Codon Models, a method for grouping sites according to similarities in the underlying process of evolution. *Mol. Biol. Evol.* 25:1995-2007.

Bielawski, J.P. and Yang, Z. 2001. The role of selection in the evolution of the DAZ gene family. *Mol. Biol. Evol.* 18:523-529.

Bielawski, J.P. and Yang, Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* 59:121-132.

Bielawski, J.P. and Yang, Z. 2005. Maximum likelihood methods for detecting adaptive protein evolution. *In* Statistical Methods in Molecular Evolution (R. Nielsen, ed.) pp. 103-124. Springer-Verlag, New York.

Bielawski, J.P., Dunn, K.A., Sabehi, G., and Béjà, O. 2004. Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc. Natl. Acad. Sci. U.S.A.* 101:14824-14829.

DeLong, E.F. and Béjà, O. 2010. The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. *PLoS Biol.* 8:e1000359.

Dutheil, J.Y., Galtier, N., Romiguier, J., Douzery, E.J., Ranwez, V., and Boussau, B. 2012. Efficient selection of branch-specific models of sequence evolution. *Mol. Biol. Evol.* 29:1861-1874.

Field, S. F., Bulina, M.Y., Kelmanson, I.V., Bielawski, J.P., and Matz, M.V. 2006. Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J. Mol. Evol.* 62:332-339.

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

Guindon, S., Rodrigo, A.G., Dyer, K.A., and Huelsenbeck, J.P. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci. U.S.A.* 101:12957-12962.

Jiggins, F.M., Hurst, G.D.D., and Yang, Z. 2002. Host-symbiont conflicts: Positive selection on the outer membrane protein of parasite but not mutualistic Rickettsiaceae. *Mol. Biol. Evol.* 19:1341-1349.

Kelley, J.L. and Swanson, W.J. 2008. Dietary change and adaptive evolution of enamelin in humans and among primates. *Genetics* 178:1595-1603.

Kosakovsky Pond, S.L. and Frost, S.D. 2005a. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.* 22:478-485.

Kosakovsky Pond, S.L. and Frost, S.D. 2005b. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22:1208-1222.

Kosakovsky Pond, S.L. and Muse, S.V. 2005. HyPhy: Hypothesis testing using phylogenies. *In* Statistical Methods in Molecular Evolution (R. Nielsen, ed.) pp. 125-181. Springer-Verlag, New York.

Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* 28:3033-3043.

Muse, S.V. and Gaut, B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* 11:715-725.

Pawitan, Y. 2001. In all likelihood: Statistical modeling and inference using likelihood. Clarendon Press, Oxford.

Rodrigue, N., Lartillot., N., and Philippe, H. 2008. Bayesian comparisons of codon substitution models. *Genetics* 180:1579-1591.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

Swofford, D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568-573.

Yang, Z. 2006. Computational Molecular Evolution. Oxford University Press, Oxford.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586-1591.

Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *TREE* 15:496-503.

Yang, Z. and dos Reis, M. 2011. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* 28:1217-1228.

Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32-43.

Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19:908-917.

Yang, Z. and Swanson, W.J. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* 19:49-57.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431-449.

Yap, V.B., Lindsay, H., Easteal, S., and Huttley, G. 2010. Estimates of the effect of natural selection on protein-coding content. *Mol. Biol. Evol.* 27:726-734.

Zhang, J., Nielsen, R., and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22:2472-2479.

Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.