# Maximum Likelihood Methods for Detecting Adaptive Protein Evolution

Joseph P. Bielawski[1] and Ziheng Yang[2]

April 19, 2004

[1] Department of Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4J1, Canada

[2] Department of Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom

# 1 Introduction

Proteins evolve; the genes encoding them undergo mutation, and the evolutionary fate of the new mutation is determined by random genetic drift as well as purifying or positive (Darwinian) selection. The ability to analyze this process was realized in the late 1970's when techniques to measure genetic variation at the sequence level were developed. Arrival of molecular sequence data also intensified the debate concerning the relative importance of neutral drift and positive selection to the process of molecular evolution (*e.g.*, Kimura and Ohta 1974). Ever since there has been considerable interest in documenting cases of molecular adaptation. Despite a spectacular increase in the amount of available nucleotide sequence data over the last three decades, the number of such well-established cases is still relatively small (*e.g.*, Endo et al. 1996; Yang and Bielawski 2000). This is largely due to the difficulty in developing powerful statistical tests for adaptive molecular evolution. Although several powerful tests for non-neutral evolution have been developed (*e.g.*, Wayne and Simonsen 1998), significant results under such tests do not necessarily indicate evolution by positive selection.

A powerful approach to detecting molecular evolution by positive selection derives from comparison of the relative rates of synonymous and nonsynonymous substitutions (*e.g.*, Miyata and Yasunaga 1980). Synonymous mutations do not change the amino acid sequence; hence their substitution rate ($d_S$) is neutral with respect to selective pressure on the protein product of a gene. Nonsynonymous mutations do change the amino acid sequence, so their substitution rate ($d_N$) is a function of selective pressure on the protein. The ratio of these rates ($\omega = d_N/d_S$) is a measure of selective pressure. For example, if nonsynonymous mutations are deleterious, purifying selection will reduce their fixation rate and $d_N/d_S$ will be less than 1, whereas if nonsynonymous mutations are advantageous they will be fixed at a higher rate than synonymous mutations, and $d_N/d_S$ will be greater than 1. A $d_N/d_S$ ratio equal to one is consistent with neutral evolution.

With the advent of genome-scale sequencing projects we can begin to study the mechanisms of innovation and divergence at a new dimension. Undoubtedly new examples of adaptive evolution will be uncovered; however, we will also be able to study the process of molecular adaptation in the context of the amount and nature of genomic change involved. Statistical tools such as maximum likelihood estimation of the $d_N/d_S$ ratio (Goldman and Yang 1994; Muse and Gaut 1994) and the likelihood ratio test for positively selected genes (*e.g.*, Nielsen and Yang 1998; Yang et al. 2000) will be valuable assets in this effort. Hence, the objective of this chapter is to provide an overview of some recent developments in statistical methods for detecting adaptive evolution as implemented in the PAML package of computer programs.

## 1.1 The PAML package of programs

PAML (for Phylogenetic Analysis by Maximum Likelihood) is a package of programs for analysis of DNA or protein sequences by using maximum likelihood methods in a phylogenetic framework (Yang 1997). The package, along with documentation and source codes, is available at the PAML web site (http://abacus.gene.ucl.ac.uk/software/paml.html). In this chapter we illustrate selected topics by analysis of example data sets. The sequence alignments, phylogenetic trees, and the control files for running the program are all available at ftp://abacus.gene.ucl.ac.uk/pub/BY2004SMME/. Readers are encouraged to retrieve and analyze the example datasets themselves as they proceed through this chapter.

The majority of analytical tools discussed here are implemented in the `codeml` program in the PAML package. Data analysis using `codeml`, and the other programs in the PAML package, are controlled by variables listed in a "control file". The control file for `codeml` is called `codeml.ctl` and is read and modified by using a text editor. Options that do not apply to a particular analysis can be deleted from a control file. Detailed descriptions of all of `codeml`'s variables are provided in the PAML documentation. Below we list a sample file, showing the important options for codon-based analysis discussed in this chapter.

```
 seqfile = seqfile.txt   * sequence data filename
treefile = tree.txt      * tree structure file name
```

```
  outfile = out.txt
  runmode = 0      * 0:user defined tree; -2:pairwise comparison
  seqtype = 1      * 1:codon models; 2: amino acid models
CodonFreq = 2      * 0:equal, 1:F1X4, 2:F3X4, 3:F61
    model = 0      * 0:one-w for all branches; 2: w's for branches
  NSsites = 0      * 0:one-rtio; 1:neutral; 2:selection; 3:discrete;
                   * 7:beta; 8:beta&w
    icode = 0      * 0:universal code
fix_kappa = 0      * 1:kappa fixed, 0:kappa to be estimated
    kappa = 2      * initial or fixed kappa
fix_omega = 0      * 1:omega fixed, 0:omega to be estimated
    omega = 5      * initial omega
```

# 2 Maximum likelihood estimation of selective pressure for pairs of sequences

## 2.1 Markov model of codon evolution

A Markov process is a simple stochastic process in which the probability of change from one state to another depends on the current state only and not on past states. Markov models have been used very successfully to describe changes between nucleotides, codons or amino acids (*e.g.*, Felsenstein 1981; Goldman and Yang 1994; Kishino et al. 1990). Advantages of a codon model include the ability to model biologically important properties of protein coding sequences such as the transition to transversion rate ratio, the $d_N/d_S$ ratio and codon usage frequencies. Since we are interested in measuring selective pressure by using the $d_N/d_S$ ratio, we will consider a Markov process which describes substitutions between the 61 sense codons within a protein coding sequence (Goldman and Yang 1994). The three stop codons are excluded because mutations to stop codons are not tolerated in a functional protein coding gene. Independence among the codon sites of a gene is assumed, and hence the substitution process can be considered one codon site at a time. For any single codon site, the model describes the instantaneous substitution rate from codon $i$ to codon $j$, $q_{ij}$. Because transitional substitutions are known to occur more often than transversions, the rate is multiplied by the $\kappa$ parameter when the change involves a transition; the $\kappa$ parameter is the transition/transversion rate ratio. Usage of codons within a gene also can be highly biased, and consequently, the rate of change from $i$ to $j$ is multiplied by the equilibrium frequency of codon $j$ ($\pi_j$). Selective constraints acting on substitutions at the amino acid level affect the rate of change when that change represents a nonsynonymous substitution. To account for this level of selective pressure the rate is multiplied by the $\omega$ parameter if the change is nonsynonymous; the $\omega$ parameter is the nonsynonymous/synonymous rate ratio ($d_N/d_S$). Note that only selection on the protein product of the gene influences $\omega$.

The substitution model is specified by the instantaneous rate matrix, $Q = \{q_{ij}\}$, where

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and j differ at two or three codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and j differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and j differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and j differ by a nonsynonymous transversion,} \\ \mu\kappa\omega\pi_j, & \text{if } i \text{ and j differ by a nonsynonymous transition.} \end{cases} \tag{1}$$

The diagonal elements of the matrix $Q$ are defined by the mathematical requirement that the row sums are equal to zero. Because separate estimation of the rate ($\mu$) and time ($t$) is not possible, the rate ($\mu$) is fixed so that the expected number of nucleotide substitutions per codon is equal to one. This scaling allows us to measure time ($t$) by the expected number of substitutions per codon, *i.e.*, genetic distance. The probability that codon $i$ is substituted by codon $j$ after time $t$ is $p_{ij}(t)$, and $P(t) = p_{ij}(t) = e^{Qt}$. The above is a description of the basic codon model of Goldman and Yang (1994). A similar model of codon substitution was proposed by Muse and Gaut (1994), which is implemented in `codeml` as well as in the program HyPhy (http://www.hyphy.org/).

## 2.2 Maximum likelihood estimation of the $d_N/d_S$ ratio

We can estimate $\omega$ by maximizing the likelihood function using data of two aligned sequences. Suppose there are $n$ codon sites in a gene, and a certain site ($h$) has codons CCC and CTC. The data at site $h$, denoted $x_h = \{CCC, CTC\}$, are related to an ancestor with codon $k$ by branch lengths $t_0$ and $t_1$ (Fig. 1a). The probability of site $h$ is

$$f(x_h) = \sum_k \pi_k p_{k,CCC}(t_0) p_{k,CTC}(t_1) = \pi_{CCC} p_{CCC,CTC}(t_0 + t_1). \quad (2)$$

Since the ancestral codon is unknown, the summation is over all 61 possible codons for $k$. Furthermore, as the substitution model is time-reversible, the root of the tree can be moved around, say, to species 1, without changing the likelihood. Thus $t_0$ and $t_1$ cannot be estimated individually, and only $t_0 + t_1 = t$ is estimated (Fig. 1b).

The log likelihood function is a sum over all codon sites in the sequence

$$\ell(t, \kappa, \omega) = \sum_{h=1}^{n} \log f(x_h). \quad (3)$$

Codon frequencies ($\pi_i$'s) can usually be estimated by using observed base or codon frequencies. The $\omega$ parameter, as well as parameters $\kappa$ and $t$, are estimated by maximizing the log likelihood function. Since an analytical solution is not possible, numerical optimization algorithms are used.

## 2.3 Empirical demonstration: pairwise estimation of the $d_N/d_S$ ratio for *GstD1*

In this section we use a simple dataset and the `codeml` program to illustrate maximum likelihood estimation of $\omega$. The dataset is *GstD*1 genes of *Drosophila melanogaster* and *D. simulans*. The alignment has 600 codons. Our first objective is to evaluate the likelihood function for a variety of fixed values for the parameter $\omega$. `Codeml` uses a hill-climbing algorithm to maximize the log likelihood function. In this case we will let `codeml` estimate $\kappa$ (`fix_kappa = 0` in the control file `codeml.ctl`) and the sequence distance $t$, but with parameter $\omega$ fixed (`fix_omega = 1`). All that remains is to run `codeml` several times, each with a different value for `omega` in the control file; the data in Fig. 2 show the results for ten different values of $\omega$. Note that the maximum likelihood value for $\omega$ appears to be roughly 0.06, which is consistent with purifying selection, and that values greater than 1 have much lower likelihood scores.

Our second objective is to allow `codeml` to use the hill-climbing algorithm to maximize the log likelihood function with respect to $\kappa$, $t$ and $\omega$. Thus we use `fix_omega = 1`, and can use any positive value for `omega`, which is used only as a starting value for the iteration. Such a run gives the estimate of $\omega$ of 0.067.

Alternatives to maximum likelihood estimates of $\omega$ are common (*e.g.*, Nei and Gojobori 1986; Ina 1996; Yang and Nielsen 2000). Those methods count the number of sites and differences and then apply a multiple-hit correction and are termed the counting methods. Most of them make simplistic assumptions about the evolutionary process, and apply *ad hoc* treatments to the data that can't be justified (Muse 1996; Yang and Nielsen 2000). Here we use the *GstD1* sequences to explore the effects of (i) ignoring the transition to transversion rate ratio (`fix_kappa = 1`; `kappa = 1`); (ii) ignoring codon usage bias (`CodonFreq = 0`); and (iii) alternative treatments of unequal codon frequencies (`CodonFreq = 2` and `CodonFreq = 3`). Note that for these data transitions are occurring at higher rates than transversions and codon frequencies are very biased, with average base frequencies of 6% (T), 50% (C), 5% (A) and 39% (G) at the third position of the codon. Thus, we expect estimates accounting for both biases will be the most reliable.

Results of our exploratory analyses (Table 1) indicate that model assumptions are very important for these data. For example, ignoring the transition to transversion ratio almost always led to underestimation of the number of synonymous sites ($S$), overestimation of $d_S$, and underestimation of $\omega$. This is because transitions at the third codon positions are more likely to be synonymous than transversions are (Li 1985). Similarly, biased codon usage implies unequal substitution rates between the codons, and ignoring it also leads to biased estimates of synonymous and nonsynonymous substitution rates. In real data analysis, codon usage bias was noted to have an even greater impact than the transition/transversion rate ratio, and is opposite to that of ignoring transition bias. This is clearly indicated by the sensitivity of $S$ to codon bias, where $S$ in this gene (45.2) is less than one third the expected value under the assumption of no codon bias ($S = 165.8$). The estimates of $\omega$ differ by as much as 4.7 fold (Table 1). Note that these two sequences differed at just 3% of sites.

For comparison we included estimates obtained from two counting methods. The method of Nei and Gojobori (1986) is similar to ML ignoring transition bias and codon bias, whereas the method of Yang and Nielsen (2000) is similar to ML accommodating transition bias and codon bias (F3x4). Note that estimation according to Nei and Gojobori (1986) was accomplished by using the `codeml` program, and according to Yang and Nielsen (2000) by using the YN00 program of PAML. What is clear from these data is that when sequence divergence is not too great, assumptions appear to matter more than methods, with ML and the counting methods giving similar results under similar assumptions. This result is consistent with simulation studies examining the performance of different estimation methods (Yang and Nielsen 2000). However, as sequence divergence increases, *ad hoc* treatment of the data can lead to serious estimation errors (Muse 1996; Dunn et al. 2001).

## 3 Phylogenetic estimation of selective pressure

Adaptive evolution is very difficult to detect using the pairwise approach to estimating the $d_N/d_S$ ratio. For example, a large-scale database survey identified less than 1% of genes (17 out of 3,595) as evolving under positive selective pressure (Endo et al. 1996). The problem with the pairwise approach is that it averages selective pressure over the entire evolutionary history separating the two lineages and over all codon sites in the sequences. In most functional genes, the majority of amino acid sites will be subject to strong purifying selection (Sharp 1997; Crandall et al. 1999), with only a small fraction of the sites potentially targeted by adaptive evolution (Golding and Dean 1998). In such cases, averaging the $d_N/d_S$ ratio over all sites will yield values much less than one, even under strong positive selective pressure at some sites. Moreover, if a gene evolved under purifying selection for most of that time, with only brief episodes of adaptive evolution, averaging over the history of two distantly related sequences would be unlikely to produce a $d_N/d_S$ ratio greater than one (*e.g.*, Bielawski and Yang 2001). Clearly, the pairwise approach has low power to detect positive selection. Power is improved if selective pressure is allowed to vary over sites or branches (*e.g.*, Yang 1998; Yang et al. 2000). However, increasing the complexity of the codon model in this way requires that likelihood is calculated for multiple sequences on a phylogeny.

## 3.1 Likelihood calculation for multiple sequences on a phylogeny

Likelihood calculation on a phylogeny (Fig. 3) is an extension of the calculation for two lineages. As in the case of two sequences, the root cannot be identified and is fixed at one of the ancestral nodes arbitrarily. For example, given an unrooted tree with four species and two ancestral codons, $k$ and $g$, the probability of observing the data at codon site $h$, $x_h = \{x_1, x_2, x_3, x_4\}$ (Fig. 3), is

$$f(x_h) = \sum_k \sum_g \left\{ \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2) p_{kg}(t_0) p_{gx_3}(t_3) p_{gx_4}(t_4) \right\}. \tag{4}$$

The quantity in the brackets is the contribution to the probability of observing the data by ancestral codons $k$ and $g$ at the two ancestral nodes. For an unrooted tree of $N$ species, with $N-2$ ancestral nodes, the data at each site will be a sum over $61^{N-2}$ possible combinations of ancestral codons. The log likelihood function is a sum over all codon sites in the alignment

$$\ell = \sum_{h=1}^{n} \log\{f(x_h)\}. \tag{5}$$

As in the two-species case, numerical optimization is used to maximize the likelihood function with respect to $\kappa, \omega$, and the $(2N-3)$ branch length parameters ($t$'s).

## 3.2 Modelling variable selective pressure among lineages

Adaptive evolution is most likely to occur in an episodic fashion. For example, functional divergence of duplicated genes (*e.g.*, Zhang et al. 1998; Schmidt et al. 1999; Bielawski and Yang 2003), colonization of a host by a parasitic organism (*e.g.*, Jiggins et al. 2002), or colonization of a new ecological niche (*e.g.*, Messier and Stewart 1997) all seem to occur at particular time points in the evolutionary history. To improve detection of episodic adaptive evolution, Yang (1998; see also

Muse and Gaut 1994) implemented models that allow for different $\omega$ parameters in different parts of a phylogeny. The simplest model, described above, assumes the same $\omega$ ratio for all branches in the phylogeny. The most general model, called the "free ratios model", specifies an independent $\omega$ ratio for each branch in a phylogeny. In the `codeml` program, users can specify an intermediate model, with independent $\omega$ parameters for different sets of branches. Modelling variable selective pressure involves a straightforward modification of the likelihood computation (Yang 1998). Consider the example tree of Fig. 4. Suppose we suspect selective pressure has changed in one part of this tree, perhaps due to positive selective pressure. To model this, we specify independent $\omega$ ratios ($\omega_0$ and $\omega_1$) for the two different sets of branches (Fig. 4). The transition probabilities for the two sets of branches are calculated from different rate matrices ($Q$) generated by using different $\omega$ ratios. Under this model (Fig. 4) the probability of observing the data at codon site $x_h$ is

$$f(x_h) = \sum_k \sum_g \pi_k p_{kx_1}(t_1;\omega_0) p_{kx_2}(t_2;\omega_0) p_{kg}(t_0;\omega_0) p_{gx_3}(t_3;\omega_1) p_{gx_4}(t_4;\omega_1). \tag{6}$$

The log likelihood function remains a sum over all sites, but is now maximized with respect to $\omega_0$ and $\omega_1$, as well as branch lengths ($t$s) and $\kappa$. $\omega$ parameters for user defined sets of branches are specified by "model = 2" in the control file and by labelling branches in the tree, as described in the PAML documentation.

### 3.3 Modelling variable selective pressures among sites

In practice, modelling variable selective pressure among sites appears to provide much greater gains in power than does modelling variable selective pressure among branches (Yang and Bielawski 2000). This is because adaptive evolution is generally restricted to a small subset of sites (*e.g.*, Crandall et al. 1999; Yang et al. 2000), and the previous model for variation over branches effectively averages over all sites. Although differences in the relative rate of nonsyonymous substitution often can be detected among branches, averaging over sites means it is unlikely that estimated $\omega$s will be greater than one. In fact, implementation of models with variable $\omega$s among codon sites (Nielsen and Yang 1998; Yang et al. 2000; Yang and Swanson 2002) has led to the detection of positive selection in many genes for which it had not previously been observed. For example, Zanotto et al. (1999) used the models of Nielsen and Yang (1998) to detect positive selection in the *nef* gene of HIV-1, a gene for which earlier studies had found no evidence for adaptive evolution (Plikat et al. 1997; da Silva and Hughes 1998).

There are two approaches to modelling variation in $\omega$ among sites: (i) use a statistical distribution to model the random variation in $\omega$ over sites; and (ii) use *a priori* knowledge of a protein's structural and functional domains to partition sites in the protein and use different $\omega$s for different partitions. Since structural and functional information is unknown for most proteins, a statistical distribution will be the most common approach. Collectively, Nielsen and Yang (1998) and Yang et al. (2000) implemented 13 such models, available in the `codeml` program. The continuous distributions are approximated by using discrete categories. In this approach codon sites are assumed to fall into $K$ classes, with the $\omega$ ratios for the site classes, and their proportions ($p$), estimated from the data. The number of classes ($K$) is fixed beforehand, and the $\omega$s and $p$s are either treated as parameters or functions of parameters of the $\omega$ distribution (Yang et al. 2000). We illustrate likelihood calculation by taking the discrete model (M3) as an example. M3 classifies codon sites into $K$ discrete classes ($i = 0, 1, 2, , K-1$), with $d_N/d_S$ ratios and proportions given as:

$$\begin{gathered} \omega_0, \omega_1, ..., \omega_{K-1}, \\ p_0, p_1, ..., p_{K-1}. \end{gathered} \tag{7}$$

Equation 4 is used to compute the conditional probability $f(x_h|\omega_i)$ of the data at a site, $h$, for each site class. Since we do not know to which class site $h$ belongs, we sum over both classes giving the unconditional probability:

$$f(x_h) = \sum_{i=0}^{K-1} p_i f(x_h|\omega_i). \tag{8}$$

In this way the unconditional probability is an average over the site classes of the $\omega$ distribution. Still assuming that the substitution process at individual codon sites is independent, the log likelihood function is a sum over all sites in the sequence:

$$\ell = \sum_{h=1}^{n} f(x_h). \tag{9}$$

The log likelihood is now maximized as a function of the parameters of the $\omega$ distribution, branch lengths ($t$) and $\kappa$.

With the second approach, we used knowledge of a protein's structural or functional domains to classify codon sites into different partitions with different $\omega$s. Since we assume site independence, the likelihood calculation is straightforward; the transition probabilities in equation 4 are computed by using the appropriate $\omega$ parameter for each codon site. By taking this approach we are effectively assuming our knowledge of the protein is without error; hence, we do not average over site classes for each site (Yang and Swanson 2002).

## 4 Detecting adaptive evolution in real datasets

Maximum likelihood estimation of selective pressure is only one part of the problem of detecting adaptive evolution in real datasets. We also need the tools to rigorously test hypotheses about the nature of selective pressure. For example, we might want to test if $d_N$ is higher than $d_S$; *i.e.*, $\omega > 1$. Fortunately, we can combine estimation of selective pressure with a formal statistical approach to hypothesis testing, the likelihood ratio test (LRT). Combined with Markov models of codon evolution, the LRT provides a very general method for testing hypotheses about protein evolution, including: (i) a test for variation in selective pressure among branches, (ii) a test for variation in selective pressure among sites; and (iii) a test for a fraction of sites evolving under positive selective pressure. In the case of a significant LRT for sites evolving under positive selection, we use Bayes or Empirical Bayes methods to identify positively selected sites in an alignment. In the following section we provide an introduction to the LRT and Bayes theorem, and provide some empirical demonstrations of their use on real data.

### 4.1 Likelihood ratio test (LRT)

The LRT is a general method for testing assumptions (model parameters) through comparison of two competing hypotheses. For our purposes, we will only consider comparisons of nested models; *i.e.*, the null hypothesis ($H_0$) is a restricted version (special case) of the alternative hypothesis ($H_1$). Note that the LRT only evaluates the differences between a pair of models, and any inadequacies shared by both models remain untested. Let $\ell_0$ be the maximum log likelihood under $H_0$ with parameters $\theta_0$, and $\ell_1$ be the maximum log likelihood under $H_1$ with parameters $\theta_1$. The log likelihood statistic is defined as twice the log likelihood difference between the two models

$$2\Delta\ell = 2(\ell_1(\hat{\theta_1}) - \ell_0(\hat{\theta_0})) \tag{10}$$

If the null hypothesis is true, $2\Delta\ell$ will be asymptotically $\chi^2$ distributed with the degree of freedom equal to the difference in the number of parameters between the two models.

Use of the $\chi^2$ approximation to the likelihood ratio statistic requires that certain conditions are met. First, the hypotheses must be nested. Second, the sample must be sufficiently large; the $\chi^2$ approximation fails when too few data are used. Third, $H_1$ may not be related to $H_0$ by fixing one or more of its parameters at the boundary of parameter space. This is called the ”boundary” problem, and the LRT statistic is not expected to follow a $\chi^2$ distribution in this case (Self and Liang 1987). When the above conditions are not met, the exact distribution can be obtained by Monte Carlo simulation (Goldman 1993; Anisimova et al. 2001), although this can be a computationally costly solution.

## 4.2 Empirical demonstration: LRT for variation in selective pressure among branches in *Ldh*

The *Ldh* gene family is an important model system for molecular evolution of isozyme multigene families (Li and Tsoi 2002). The paralogous copies of lactate dehydrogenase (*Ldh*) genes found in mammals originated from a duplication near the origin of vertebrates (*Ldh-A* and *Ldh-B*), and a later duplication near the origin of mammals (Fig. 5; *Ldh-A* and *Ldh-C*). Li and Tsoi (2002) found that the rate of evolution had increased in mammalian *Ldh-C* sometime following the second duplication event. An unresolved question about this gene family is whether the increased rate of *Ldh-C* reflects (i) a burst of positive selection for functional divergence following the duplication event, (ii) a long term change in selective pressure or (iii) simply an increase in the underlying mutation rate of *Ldh-C*. In the following we use the LRT for variable $\omega$ ratios among branches to test these evolutionary scenarios.

The null hypothesis ($H_0$) is that the rate increase in *Ldh-C* is simply due to an underlying increase in the mutation rate. If selective pressure was constant and the mutation rate increased, the relative fixation rates of synonymous and non-synonymous mutations ($\omega$) would remain constant over the phylogeny, but the overall rate of evolution would increase in *Ldh-C*. One alternative to this scenario is that the rate increase in *Ldh-C* was due to a burst of positive selection following gene duplication ($H_1$). A formal test for variation in selective pressure among sites may be formulated as follows:

$H_0$: $\omega$ is identical across all branches of the *Ldh* phylogeny.

$H_1$: $\omega$ is variable, being greater than 1 in branch C0 of Fig. 5.

Because $H_1$ can be transformed into $H_0$ by restricting $\omega_{C0}$ to be equal to the $\omega$ ratios for the other branches, we can use the LRT. The estimate of $\omega$ under the null hypothesis, as an average over the phylogeny in Fig. 5, was 0.14, indicating that evolution of *Ldh-A* and *Ldh-C* was dominated by purifying selection. The LRT suggests that selective pressure in *Ldh-C* immediately following gene duplication (0.19) was not significantly different than the average over the other branches (Table 2). Hence, we found no evidence for functional divergence of *Ldh-A* and *Ldh-C* by positive selection. It should be noted that if functional divergence of *Ldh-A* and *Ldh-C* evolved by positive selection for just one or a few amino acid changes, we would not observe a large difference in $\omega$ ratios among branches.

Using the same approach, we tested a second alternative hypothesis where the rate increase in *Ldh-C* was due to an increase in the nonsynonymous substitution rate over all lineages of the *Ldh-C* clade (see $H_2$ in Fig$\dot{5}$). In this case the LRT was highly significant, and the parameter estimates for the *Ldh-C* clade indicated an increase in the relative rate of nonsynonymous substitution by a factor of 3 (Table 2). Lastly, we tested the hypothesis that selective pressure differed in both *Ldh-A* and *Ldh-C* following gene duplication (see $H_3$ in Fig$\dot{5}$), and results of this test were not significant

(Table 2). Collectively, these findings suggest selective pressure and mutation rates in *Ldh-A* were relatively unchanged by the duplication event, whereas the nonsynonymous rate increased in *Ldh-C* following the duplication event, as compared with *Ldh-A*.

## 4.3 Empirical demonstration: positive selection in the *nef* gene in the human HIV-2 genome

The role of the *nef* gene in differing phenotypes of HIV-1 infection has been well studied, including identification of sites evolving under positive selective pressure (Zanotto et al. 1999). The *nef* gene in HIV-2 has received less attention, presumably because HIV-2 is associated with reduced virulence and pathogenicity relative to HIV-1. Padua et al. (2003) sequenced 44 *nef* alleles from a study population of 37 HIV-2 infected people living in Lisbon, Portugal. They found that nucleotide variation in the *nef* gene, rather than gross structural change, was potentially correlated with HIV-2 pathogenesis. In order to determine if the *nef* gene might also be evolving under positive selective pressure in HIV-2, we analysed those same data here with models of variable $\omega$ ratios among sites (Yang et al. 2000).

Following the recommendation of Yang et al. (2000) and Anisimova et al. (2001) we consider the following models: M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta), and M8 (beta&$\omega$). Models M0 and M3 were described above. M1 (neutral) specifies two classes of sites; conserved sites with $\omega = 0$, and neutral sites with $\omega = 1$. M2 (selection) is an extension of M1 (neutral), adding a third $\omega$ class that is free to take a value $> 1$. Version 3.14 of `paml/codeml` introduces a slight variation to models M1 (neutral) and M2 (selection), in that $\omega_0 < 1$ is estimated from the data rather than being fixed at 0. Those are referred to as models M1a and M2a, also used here. Under model M7 (beta), $\omega$ varies among sites according to a beta distribution with parameters $p$ and $q$. The beta distribution is restricted to the interval (0, 1); thus, M1 (neutral), M1a (NearlyNeutral) and M7 (beta) assume no positive selection. M8 (beta&$\omega$) adds a discrete $\omega$ class to the beta distribution that is free to take a value $> 1$. Under M8 (beta&$\omega$) a proportion of sites $p_0$ is drawn from a beta distribution, with the remainder ($p_1 = 1 - p_0$) having the $\omega$ ratio of the added site class. We specified $K = 3$ discrete classes of sites under M3 (discrete), and $K = 10$ under M7 (beta) and M8 (beta&$\omega$). We use an LRT comparing M0 (one-ratio) with M3 (discrete) to test for variable selective pressure among sites, and three LRTs to test for sites evolving by positive selection, comparing (i) M1 (neutral) against M2 (selection), (ii) M1a (NearlyNeutral) and M2a (PositiveSelection), and (iii) M7 (beta) against M8 (beta&$\omega$).

Maximum likelihood estimates of parameters and likelihood scores for the *nef* gene are presented in Table 3. Averaging selective pressure over sites and branches as in M0 (one-ratio) yielded an estimated $\omega$ of 0.50, a result consistent with purifying selection. The LRT comparing M0 (one-ratio) against M3 (discrete) is highly significant ($2\Delta\ell = 1087.2$, df = 4, $P < 0.01$), indicating that the selective pressure is highly variable among sites. Estimates of $\omega$ under models that can allow for sites under positive selection (M2, M2a, M3, M8) indicated a fraction of sites evolving under positive selective pressure (Table 3). To formally test for the presence of sites evolving by positive selection, we conducted LRTs comparing M1 and M2, M1a and M2a, and M7 and M8. All those LRTs were highly significant; for example, the test stastic for comparing M1 (neutral) M2 (selection) is $2\Delta\ell$ = 223.58, with $P < 0.01$, df = 2. These findings suggest that about 12% of sites in the *nef* gene of HIV-2 are evolving under positive selective pressure, with $\omega$ between 2 and 3. It is clear from Table 3 that this mode of evolution would not have been detected if $\omega$ were measured simply as an average over all sites of *nef*.

Models M2 (selection) and M8 (beta&$\omega$) are known to have multiple local optima in some datasets, often with $\omega_2$ under M2 or $\omega$ under M8 to be $< 1$ on one peak and $> 1$ on another peak. Thus it is important to run these models multiple times with different starting values (especially different $\omega$'s) and then select the set of estimates corresponding to the highest peak. Indeed, the *nef* dataset illustrates this issue. By using different initial $\omega$'s, both the global and local optima can be found.

## 4.4 Bayesian identification of sites evolving under positive Darwinian selection

Under the approach described in this chapter, a gene is considered to have evolved under positive selective pressure if (i) the LRT is significant, and (ii) at least one of the ML estimates of $\omega > 1$. Given these conditions are satisfied, we have evidence for sites under positive selection, but no information about which sites they are. Hence, the empirical Bayes approach is used to predict them (Nielsen and Yang 1998; Yang et al. 2000). To do this we compute, in turn, the posterior probability of a site under each $\omega$ site class of a model. Sites with high posterior probabilities under the class with $\omega > 1$ are considered likely to have evolved under positive selective pressure.

Say we have a model of heterogeneous $\omega$ ratios, with $K$ site classes ($i = 0, 1, 2, , K-1$). The $\omega$ ratios and proportions are $\omega_0, \omega_1, ..., \omega_{K-1}$ and $p_0, p_1, ..., p_{K-1}$, with the proportions, $p_i$, used as the prior probabilities. The posterior probability that a site with data $x_h$ is from site class $i$ is

$$P(\omega|x_h) = \frac{P(x_h|\omega_i)p_i}{P(x_h)} = \frac{P(x_h|\omega_i)p_i}{\sum_{j=0}^{K-1} P(x_h|\omega_j)p_j}. \tag{11}$$

Because the parameters used in the above equation to calculate the posterior probability are estimated by ML ($\omega_i$ and $p_i$), the approach is called empirical Bayes. By using the ML parameters in this way we ignore their sampling errors. The posterior probabilities will be sensitive to these parameter estimates, meaning that the reliability of this approach will be poor when the parameter estimates are poor, such as in small datasets or when obtained from a local optimum.

Because the *nef* dataset above is quite large, the parameter estimates are expected to be reliable (Anisimova et al. 2002). Consistent with this, ML estimates of the strength and proportion of positively selected sites in *nef* are consistent among M2, M3 and M8 (Table 3). Fig. 6 shows the posterior probabilities for the $K = 3$ site classes at each site of *nef* under model M3. Twenty-four sites were identified as having very high posterior probability ($P > 0.95$) of evolving under positive selection (site class with $\omega > 1$). Interestingly none of these sites matched the two variable sites in a proline-rich motif that is strongly associated with an asymptomatic disease profile (Padua et al. 2003). In fact, only four of the 24 sites were found in regions of *nef* considered important for function. Disruption of the important *nef* regions is associated with reduced pathogenicity in HIV-2 infected individuals (Switzer et al. 1998; Padua et al. 2003). Our results suggest that selective pressure at such sites is fundamentally different than selection acting at the 24 positive selection sites predicted using the Bayes theorem. To be identified with such high posterior probabilities, the predicted sites must have been evolving under long-term positive selective pressure, suggesting that they are more likely subjected to immune-driven diversifying selection at epitopes (Zanotto et al. 1999; Yang et al. In press).

## 5 Power, accuracy and robustness

The boundary problem mentioned above applies to the LRT for variable selective pressure among sites and the LRT for positive selection at a fraction of sites (Anisimova et al. 2001). The problem arises in the former because the null (M0) is equivalent to M3 ($K = 3$) with 2 of the five parameters ($p_0$ and $p_1$) fixed to 0, which is at the boundary of parameter space. In comparisons of M1 with M2, M1a with M2a, and M7 with M8, the null is equivalent to the alternative with a proportion parameter ($p$) fixed to 0. Therefore, the $\chi^2$ approximation is not expected to hold. Anisimova et al. (2001) used computer simulation to investigate the effect of the boundary problem on the power and accuracy of the LRT. Use of the $\chi^2$ makes the LRT conservative, meaning that the false positive rate will be less than predicted by the specified significance level of the test (Anisimova et al. 2001). Nevertheless, the test was found to be powerful, sometimes reaching 100% in datasets consisting of 17 sequences. Power was low for highly similar and highly divergent sequences, but was modulated by the length of the sequence and the strength of positive selection. Note that simulation studies, both with and without the boundary problem, indicate that the sample size requirements for the $\chi^2$ approximation are met with relatively short sequences; in some cases as few as 50 codons (Anisimova et al. 2001).

Bayesian prediction of sites evolving under positive selection is a more difficult task than ML parameter estimation or likelihood ratio testing. The difficulty arises because the posterior probabilities depend on the (i) information contained at just a single site in the dataset, and (ii) the quality of the ML parameter estimates. Hence, a second study was conducted by Anisimova et al. (2002) to examine the power and accuracy of the Bayesian site identification. The authors made the following generalizations: (i) prediction of positively selected sites is not practical from just a few highly similar sequences; (ii) the most effective method of improving accuracy is to increase the number of lineages; and (iii) site prediction is sensitive to sampling errors in parameter estimates and to the assumed $\omega$ distribution.

Robustness refers to the stability of results to changes in the model assumptions. The LRT for positive selection is generally robust to the assumed distribution of $\omega$ over sites (Anisimova et al. 2001). However, as the LRT of M0 with M3 is a test of variable selective pressure among sites, caution must be exercised when only the M0-M3 comparison suggests positive selection. One possiblity is to use M2, which tends to be more conservative than the other models (Anisimova et al. 2002). Another approach is to select the subset of sites which are robust to the $\omega$ distribution (Anisimova et al. 2002; Yang et al. 2003). A third approach is to select sites which are robust to sampling lineages (Yang et al. 2003). We believe that sensitivity analysis is a very important part of detecting positive selection, and we make the following recommendations: (i) multiple models should be used, (ii) care should be taken to identify and discard results obtained from local optima, and (iii) assumptions such as the $\omega$ distribution or the method of correcting for biased codon frequencies should be evaluated relative to their effects on ML parameter estimation and Bayesian site prediction.

All codon models discussed above ignore the effect of the physiochemical property of the amino acid being substituted. For example, all amino acid substitutions at a positively selected site are assumed to be advantageous, with $\omega > 1$. The assumption appears to be unrealistic; one can imagine that there might be a set of amino acid substitutions that are forbidden at a site because of physiochemical constraints, even though the site is subject to strong positive selection. Another limitation is that these methods are very conservative, only indicating positive selection when the estimate of $\omega$ is $> 1$. In cases where only one or a few amino acid substitutions result in a substantial change in phenotype, the methods will have little or no power because $\omega$ will be $< 1$. Another important limitation is the assumption of a single underlying phylogeny. When recombination has occurred, no single phylogeny will fit all sites of the data. A recent simulation study (Anisimova et al. 2003) found that the LRT is robust to low levels of recombination, but can have a seriously high type I error rate when recombination is frequent. Interestingly, Bayesian prediction of positively selected sites was less affected by recombination than was the LRT. In summary, no matter how robust the results, they must be interpreted with these limitations in mind.

**Acknowledgment.** We thank an anonymous referee for many constructive comments. This work is supported by a grant from BBSRC to Z.Y.

## 6 Refrences

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

ANISIMOVA, M., J. P. BIELAWSKI, and Z. YANG. 2002. Accuracy and power of Bayesian prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19:950-958.

ANISIMOVA, M., R. NIELSEN, and Z. YANG. 2003. Effect of recombination on the accuracy of likelihood methods for detecting positive selection at amino acid sites. Genetics 164:1229-1236.

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

BIELAWSKI, J. P. and Z. YANG. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J. Struct. Funct. Genomics 3:201-212.

CRANDALL, K. A., C. R. KELSEY, H. IMANICHI, H. C. LANE, and N. P. SALZMAN. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. 16:372-382.

DA SILVA, J. and A. L. HUGHES. 1998. Conservation of cytotoxic T lymphocytes (CTL) epitopes as a host strategy to constrain parasitic adaptation: evidence from the *nef* gene of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 15: 1259-1268.

DUNN, K. D., J. P. BIELAWSKI, and Z. YANG. 2001. Rates and patterns of synonymous substitutions in *Drosophila*: implications for translational selection. Genetics 157:295-305.

ENDO, T. K., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13: 685-690.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

GOLDING, G. B., and A. M. DEAN. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. 15: 355-369.

GOLDMAN, N. 1993. Statistical tests of DNA substitution models. J. Mol. Evol. 36:182-198.

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

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular using clock mitochondrial DNA. J. Mol. Evol. 22:160-174.

HUGHES, A. L. and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.

INA, Y. 1996. Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. J. Genet. 75:91-115.

JIGGINS, F.M., G.D.D. HURST, and Z. YANG. 2002. Host-symbiont conflicts: positive selection on the outer membrane protein of parasite but not mutualistic Rickettsiaceae. Mol. Biol. Evol. 19: 1341-1349.

KIMURA, M., and T. OHTA. 1974. On some principles governing molecular evolution. Proc. Nat. Acad. Sci., USA 71: 2848-2852.

KISHINO, H., T., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151-160.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2:150-174.

LI, Y.-J. and C.-M. TSOI. 2002. Phylogenetic analysis of vertebrate lactate dehydrogenase (LDH) multigene families. J. Mol. Evol. 54:614-624.

MESSIER, W. and C.-B. STEWART 1997. Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J. Mol. Evol. 16:23-36.

MUSE, S. V. 1996. Estimating synonymous and non-synonymous substitution rates. Mol. Biol. Evol. 13:105-114.

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

NEI, M. and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

NIELSEN, R. and Z. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

PADUA, E., A. JENKINS, S. BROWN, J. BOOTMAN, M. T. PAIXAO, N. ALMOND, and N. BERRY. 2003. Natural variation of the *nef* gene in human immunodeficiency virus type 2 infections in Portugal. J. Gen. Virol. 84:1287-1299.

PLIKAT, U., K NIESELT-STRUWE, and A. MEYERHANS. 1997. Genetic drift can determine short-term human immunodeficiency virus type 1 *nef* quasispecies evolution in vivo. J. Virol. 71:4233-4240.

SCHMIDT, T. R., M. GOODMAN and L. I. GROSSMAN. 1999. Molecular evolution of the COX7A gene family in primates. Mol. Biol. Evol. 16:619-626.

SELF, S. and K. -Y. LIANG. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J. Am. Stat. Assoc. 82:605-610.

SHARP, P. M. 1997. In search of molecular Darwinism. Nature 385:111-112.

SWITZER, W. M., S. WIKTOR, and 8 CO-AUTHORS. 1998. Evidence of *nef* truncation in human immunodeficiency virus type 2 infection. J. Infect. Dis. 177:65-71.

WAYNE, M. and K. SIMONSEN 1998. Statistical tests of neutrality in an age of weak selection. TREE 13:236-240.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306-314.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Appl. Biosci. 13:555-556.

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

YANG, Z. and J. P. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. TREE 15:496-503.

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

YANG, Z. and W. J. SWANSON. 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., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selective pressure at amino acid sites. Genetics 155:431-449.

YANG, W., J. P. BIELAWSKI, and Z. YANG. 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. In press.

ZANOTTO, P. M., E. G. KALLIS, R. F. SOUZA, and E. C. HOLMES. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153:1077-1089.

ZHANG, J. and H. F. ROSENBERG. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc. Natl. Acad. Sci. USA 99:5486-5491.

ZHANG, J., H. F. ROSENBERG, and M. NEI 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA 95:3708-3713.

## Table 1

**Estimation of $d_S$ and $d_N$ between *Drosophila melanogaster* and *D. simulans GstD1* genes**

| Method | $\kappa$ | $S$ | $N$ | $d_S$ | $d_N$ | $\omega$ | $\ell$ |
|---|---|---|---|---|---|---|---|
| ML methods | | | | | | | |
| Fequal, $\kappa$ = 1 | 1 | 152.9 | 447.1 | 0.0776 | 0.0213 | 0.274 | -927.18 |
| Fequal, $\kappa$ estimated | 1.88 | 165.8 | 434.2 | 0.0221 | 0.0691 | 0.320 | -926.28 |
| F3x4, $\kappa$ = 1 | 1 | 70.6 | 529.4 | 0.1605 | 0.0189 | 0.118 | -844.51 |
| F3x4, $\kappa$ estimated | 2.71 | 73.4 | 526.6 | 0.1526 | 0.0193 | 0.127 | -842.21 |
| F61, $\kappa$ = 1 | 1 | 40.5 | 559.5 | 0.3198 | 0.0201 | 0.063 | -758.55 |
| F61, $\kappa$ estimated | 2.53 | 45.2 | 554.8 | 0.3041 | 0.0204 | 0.067 | -756.57 |
| Counting methods | | | | | | | |
| Nei and Gojobori | 1 | 141.6 | 458.4 | 0.0750 | 0.0220 | 0.288 | |
| Yang and Nielsen (F3x4) | 3.28 | 76.6 | 523.5 | 0.1499 | 0.0190 | 0.127 | |

**Table 2 Parameter estimates under models of variable $\omega$ ratios among lineages for the *Ldh-A* and *Ldh-C* gene family**

| Models | $w_{A0}$ | $w_{A1}$ | $w_{C1}$ | $w_{C0}$ | $\ell$ |
|---|---|---|---|---|---|
| $H_0 : w_{A0} = w_{A1} = w_{C1} = w_{C0}$ | 0.14 | $= w_{A0}$ | $= w_{A0}$ | $= w_{A0}$ | -6018.63 |
| $H_1 : w_{A0} = w_{A1} = w_{C1} \neq w_{C0}$ | 0.13 | $= w_{A0}$ | $= w_{A0}$ | 0.19 | -6017.57 |
| $H_2 : w_{A0} = w_{A1} \neq w_{C1} = w_{C0}$ | 0.07 | $= w_{A0}$ | 0.24 | $= w_{A1}$ | -5985.63 |
| $H_3 : w_{A0} \neq w_{A1} \neq w_{C1} = w_{C0}$ | 0.09 | 0.06 | 0.24 | $= w_{A1}$ | -5984.11 |

Note: The topology and branch specific $\omega$ ratios are presented in Fig$\dot{5}$. The d.f. is 1 for the comparisons of H0 vs. H1, H0 vs. H2, and H2 vs. H3.

**Table 3 Parameter estimates and likelihood scores under models of variable $\omega$ ratios among sites for HIV-2 *nef* genes**

| Model | $d_N/d_S$ | Parameter estimates | PSS | $\ell$ |
|---|---|---|---|---|
| M0: one-ratio (1) | 0.51 | $\omega = 0.505$ | none | -9775.77 |
| M3: discrete (5) | 0.63 | $p_0 = 0.48, p_1 = 0.39, (p_2 = 0.13)$ | 31 (24) | -9232.18 |
| | | $\omega_0 = 0.03, \omega_1 = 0.74, \omega_2 = 2.50$ | | |
| M1: neutral (1) | 0.63 | $p_0 = 0.37, (p_1 = 0.63)$ | not allowed | -9428.75 |
| | | $(\omega_0 = 0), (\omega_1 = 1)$ | | |
| M2: selection (3) | 0.93 | $p_0 = 0.37, p_1 = 0.51, (p_2 = 0.12)$ | 30 (22) | -9392.96 |
| | | $(\omega_0 = 0), (\omega_1 = 1), \omega_2 = 3.48$ | | |
| M1a: NearlyNeutral (2) | 0.48 | $p_0 = 0.55, (p_1 = 0.45)$ | not allowed | -9315.53 |
| | | $(\omega_0 = 0.06), (\omega_1 = 1)$ | | |
| M2a: PositiveSelection (4) | 0.73 | $p_0 = 0.51, p_1 = 0.38, (p_2 = 0.11)$ | 26 (15) | -9241.33 |
| | | $(\omega_0 = 0.05), (\omega_1 = 1), \omega_2 = 3.00$ | | |
| M7: beta (2) | 0.42 | $p = 0.18, q = 0.25$ | not allowed | -9292.53 |
| M8: beta&$\omega$ (4) | 0.62 | $p_0 = 0.89, (p_1 = 0.11)$ | 27 (15) | -9224.31 |
| | | $p = 0.20, q = 0.33, \omega = 2.62$ | | |

Note. The number after the model code, in parentheses, is the number of free parameters in the $\omega$ distribution. The $d_N/d_S$ ratio is an average over all sites in the HIV-2 *nef* gene alignment. Parameters in parentheses are not free parameters and are presented for clarity. PSS is the number of positive selected sites, inferred with at the 50% (95%) posterior probability cutoff.

# Figure legends

Figure 1. Rooted (a) and (b) unrooted trees for a pair of sequences. Under reversible codon models, the root is unidentifiable; hence, only the sum of the branch lengths, $t = t_0 + t_1$, is estimable.

Figure 2. Log likelihood as a function of the $\omega$ parameter for a pair of *GstD1* genes from the *Drosophila melanogaster* and *D. simulans*. The maximum likelihood estimate of $\omega$ is the value that maximizes the likelihood function. Since an analytical solution is not possible, the `codeml` program uses a numerical hill-climbing algorithm to maximize l. For these data the maximum likelihood estimate of $\omega$ is 0.067, with a maximum likelihood of -756.57.

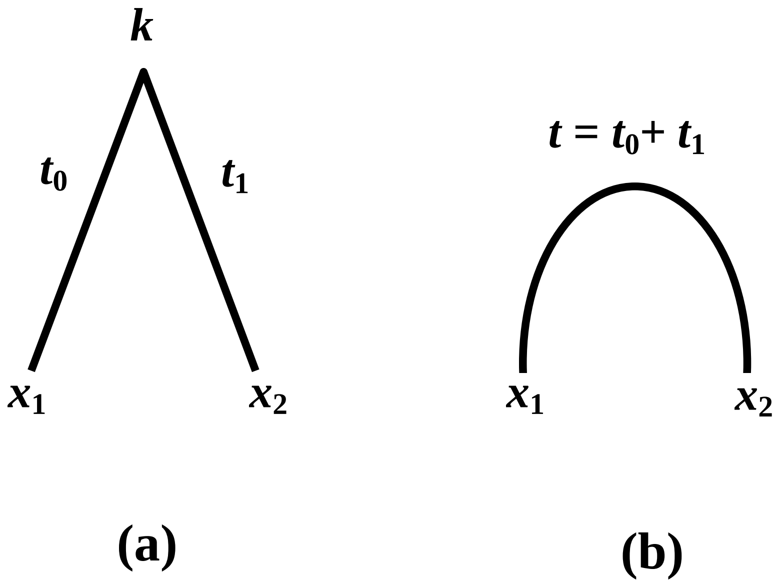
Figure 3. An unrooted phylogeny for four sequences. As in the case of two sequences, the root cannot be identified. For the purpose of likelihood calculation the root is fixed at one of the ancestral nodes arbitrarily, and $t_0, t_1, t_2, t_3$, and $t_4$ are estimable parameters in the model.

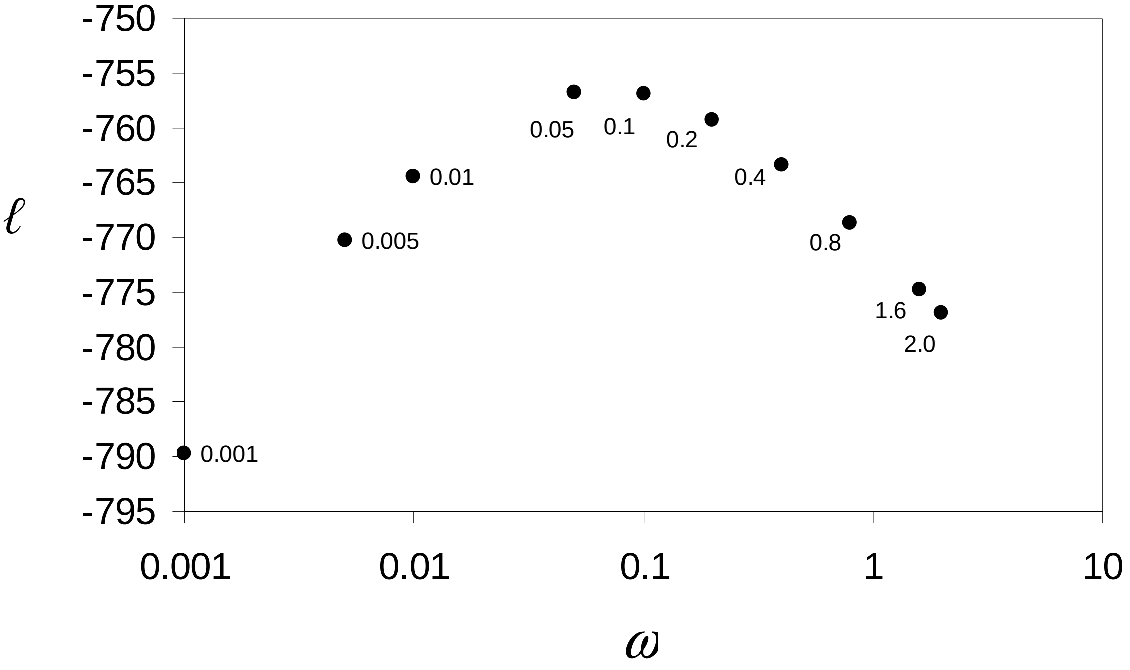
Figure 4. Four taxon phylogeny with variable $\omega$ ratios among its branches. The likelihood of this tree is calculated according to Yang (1998), where the two independent $\omega$ ratios ($\omega_0$ and $\omega_1$) are used to calculate rate matrices ($Q$) and transition probabilities for the different branches.

Figure 5. A phylogenetic tree for the *Ldh-A* and *Ldh-C* gene family. The tree was obtained by a neighbour-joining analysis of a codon sequence alignment under the HKY85 substitution model (Haegawa et al. 1985) combined with a gamma model of among sites rate variation (Yang 1994). Branch lengths are not to scale. The *Gallus* (chicken) and *Sceloporus* (lizard) *Ldh-A* sequences are pro-orthologs, as they predate the gene duplication event. The tree is rooted with the pro-orthologous sequences for convenience; all analyses were conducted by using the unrooted topology. The one ratio model ($H_0$) assumes uniform selective pressure over all branches. $H_1$ is based on the notion of a burst of positive selection in *Ldh-C* following the gene duplication event; hence the assumption of one $\omega$ for branch $C0$ and another for all other branches. $H_2$ is based on the notion of increased nonsynonymous substitution in all *Ldh-C* lineages following gene duplication; hence the assumption of one $\omega$ for the *Ldh-C* branches ($\omega_{C0} = \omega_{C1}$) and another for the *Ldh-A* branches ($\omega_{A0} = \omega_{A1}$). $H_3$ is based on the notion that selective pressure changed in both *Ldh-C* and *Ldh-A*

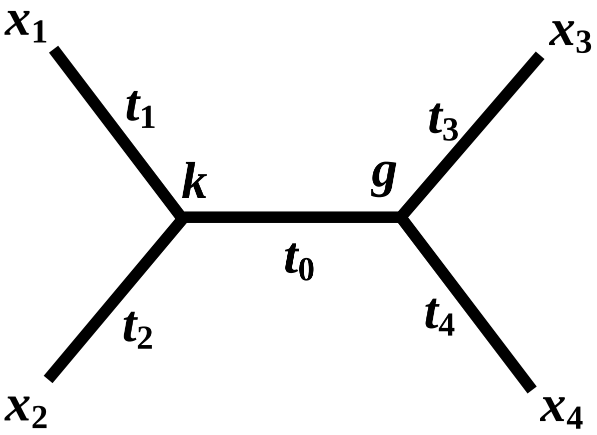
following gene duplication, as compared with the pro-orthologous sequences; hence, one $\omega$ for the *Ldh-C* branches ($\omega_{C0} = \omega_{C1}$), one $\omega$ for the post-duplication *Ldh-A* branches ($\omega_{A1}$), and one $\omega$ for the pro-orthologous branches ($\omega_{A0}$).

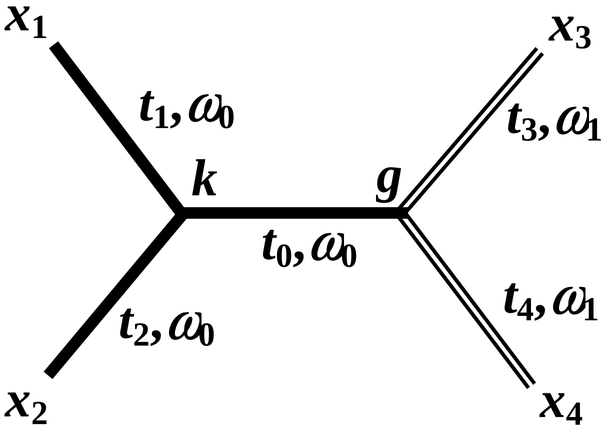
Figure 6. Posterior probabilities for sites classes under M3 ($K = 3$) along the HIV-2 *nef* gene alignment.

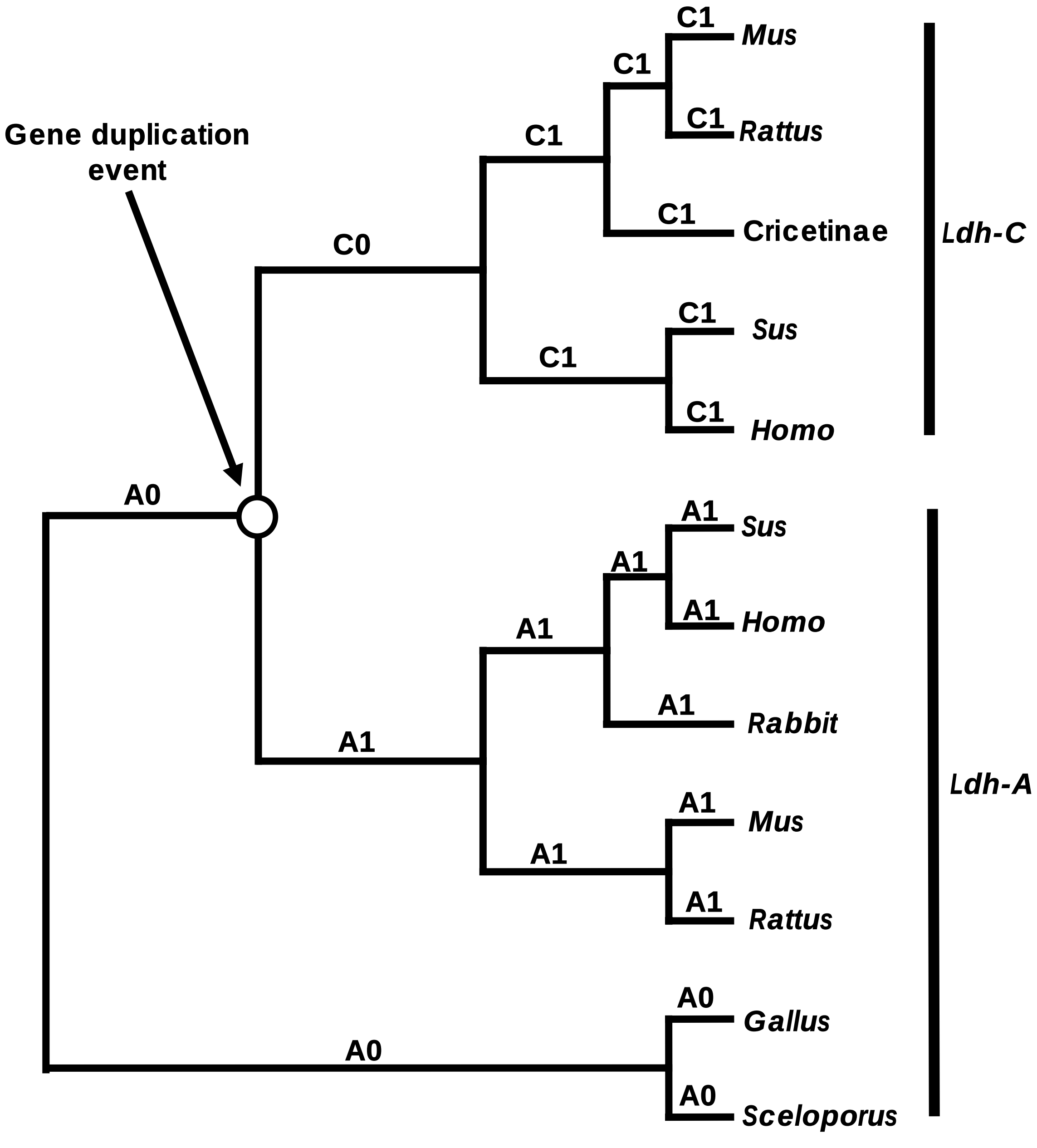

k
t0
t1
x1
x2
(a)
t = t0+ t1
x1
x2
(b)



ℓ
-750
-755
-760
-765
-770
-775
-780
-785
-790
-795
0.001
0.01
0.1
1
10
ω
0.05
0.1
0.2
0.01
0.4
0.005
0.8
1.6
2.0
0.001


$x_1$
$t_1$
$k$
$g$
$t_3$
$x_3$
$t_0$
$t_2$
$t_4$
$x_2$
$x_4$


$x_1$
$t_1,\omega_0$
$k$
$g$
$x_3$
$t_3,\omega_1$
$t_0,\omega_0$
$t_4,\omega_1$
$t_2,\omega_0$
$x_2$
$x_4$



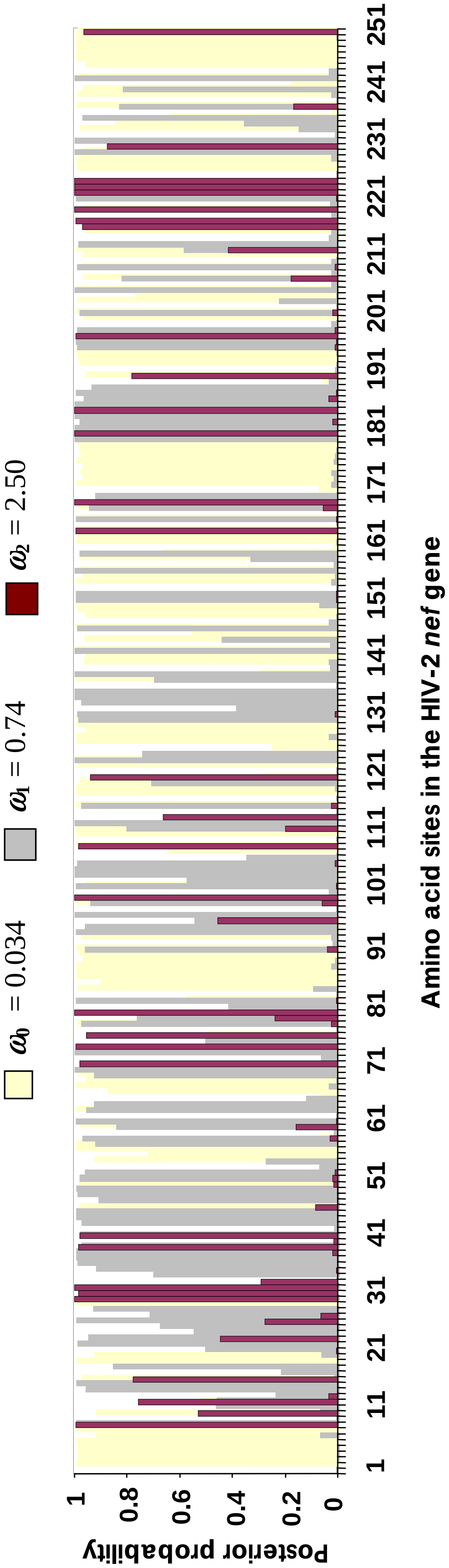
$$H_0: \omega_{A0} = \omega_{A1} = \omega_{C1} = \omega_{C0}$$

$$H_1: \omega_{A0} = \omega_{A1} = \omega_{C1} \neq \omega_{C0}$$

$$H_2: \omega_{A0} = \omega_{A1} \neq \omega_{C1} = \omega_{C0}$$

$$H_3: \omega_{A0} \neq \omega_{A1} \neq \omega_{C1} = \omega_{C0}$$

1.


$\omega_0$ = 0.034
$\omega_1$ = 0.74
$\omega_2$ = 2.50
Posterior probability
0
0.2
0.4
0.6
0.8
1
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
Amino acid sites in the HIV-2 *nef* gene