5 Maximum Likelihood Methods for Detectin Adaptive Protein Evolution

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

[1] Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4
Canada, j.bielawski@dal.ca
[2] Department of Biology, University College London, Gower Street, London
WC1E 6BT, United Kingdom, z.yang@ucl.ac.uk

5.1 Introduction

Proteins evolve; the genes encoding them undergo mutation, and the
tionary fate of the new mutation is determined by random genetic d
well as purifying or positive (Darwinian) selection. The ability to analy
process was realized in the late 1970s when techniques to measure g
variation at the sequence level were developed. The arrival of molecu
quence data also intensified the debate concerning the relative importa
neutral drift and positive selection to the process of molecular evolutio
Ever since, there has been considerable interest in documenting cases o
cular adaptation. Despite a spectacular increase in the amount of av
nucleotide sequence data since the 1970s, the number of such well-estab
cases is still relatively small [9, 38]. This is largely due to the difficulty
veloping powerful statistical tests for adaptive molecular evolution. Alt
several powerful tests for nonneutral evolution have been developed [3
nificant results under such tests do not necessarily indicate evolution b
itive selection.

A powerful approach to detecting molecular evolution by positive se
derives from comparison of the relative rates of synonymous and nons
mous substitutions [22]. Synonymous mutations do not change the
acid sequence; hence their substitution rate (d_S) is neutral with respect
lective pressure on the protein product of a gene. Nonsynonymous mut
do change the amino acid sequence, so their substitution rate (d_N) is a
tion of selective pressure on the protein. The ratio of these rates ($\omega = d$
is a measure of selective pressure. For example, if nonsynonymous mut
are deleterious, purifying selection will reduce their fixation rate and
will be less than 1, whereas if nonsynonymous mutations are advanta
they will be fixed at a higher rate than synonymous mutations, and
will be greater than 1. A d_N/d_S ratio equal to one is consistent with n
evolution.

doubtedly, new examples of adaptive evolution will be uncovered; howe
will also be able to study the process of molecular adaptation in the con
the amount and nature of genomic change involved. Statistical tools s
maximum likelihood estimation of the d_N/d_S ratio [13, 24] and the like
ratio test for positively selected genes [26, 34] will be valuable assets
effort. Hence, the objective of this chapter is to provide an overview o
recent developments in statistical methods for detecting adaptive evolu
implemented in the PAML package of computer programs.

5.1.1 The PAML Package of Programs

PAML (for Phylogenetic Analysis by Maximum Likelihood) is a p
of programs for analysis of DNA or protein sequences by using ma
likelihood methods in a phylogenetic framework [36]. The package,
with documentation and source codes, is available at the PAML We
(http://abacus.gene.ucl.ac.uk/software/paml.html). In this chapter, w
trate selected topics by analysis of example datasets. The sequence
ments, phylogenetic trees, and the control files for running the progra
all available at ftp://abacus.gene.ucl.ac.uk/pub/BY2004SMME/. Read
encouraged to retrieve and analyze the example datasets themselves a
proceed through this chapter.

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

```
  seqfile = seqfile.txt   * sequence data filename
 treefile = tree.txt      * tree structure filename
  outfile = out.txt
  runmode = 0      * 0:user defined tree; -2:pairwise compari
  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 bra
  NSsites = 0      * 0:one-rtio; 1:neutral; 2:selection; 3:di
                   * 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
```

5.2.1 Markov Model of Codon Evolution

A Markov process is a simple stochastic process in which the probab change from one state to another depends on the current state only a on past states. Markov models have been used very successfully to de changes between nucleotides, codons, or amino acids [10, 18, 13]. Adva of a codon model include the ability to model biologically important erties of protein-coding sequences such as the transition to transversio ratio, the d_N/d_S ratio, and codon usage frequencies. Since we are inte in measuring selective pressure by using the d_N/d_S ratio, we will co a Markov process that describes substitutions between the 61 sense c within a protein- coding sequence [13]. The three stop codons are ex because mutations to stop codons are not tolerated in a functional p coding gene. Independence among the codon sites of a gene is assume hence the substitution process can be considered one codon site at a For any single codon site, the model describes the instantaneous su tion rate from codon i to codon j, q_{ij}. Because transitional substitutio known to occur more often than transversions, the rate is multiplied κ parameter when the change involves a transition; the κ parameter transition/transversion rate ratio. Use of codons within a gene also highly biased, and consequently the rate of change from i to j is mul by the equilibrium frequency of codon j (π_j). Selective constraints on substitutions at the amino acid level affect the rate of change whe change represents a nonsynonymous substitution. To account for thi of selective pressure, the rate is multiplied by the ω parameter if the is nonsynonymous; the ω parameter is the nonsynonymous/synonymou ratio (d_N/d_S). Note that only selection on the protein product of th influences ω.

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

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

The diagonal elements of the matrix Q are defined by the mathem requirement that the row sums be equal to zero. Because separate estim of the rate (μ) and time (t) is not possible, the rate (μ) is fixed so th expected number of nucleotide substitutions per codon is equal to one scaling allows us to measure time (t) by the expected number of substit

scription of the basic codon model of Goldman and Yang [13]. A similar
of codon substitution was proposed by Muse and Gaut [24] and is implem
in `codeml` as well as in the program HyPhy (http://www.hyphy.org/).

5.2.2 Maximum Likelihood Estimation of the d_N/d_S Ratio

We can estimate ω by maximizing the likelihood function using d
two aligned sequences. Suppose there are n codon sites in a gene,
certain site (h) has codons CCC and CTC. The data at site h, d
$x_h = \{CCC, CTC\}$, are related to an ancestor with codon k by branch l
t_0 and t_1 (Figure 5.1(a)). The probability of site h is

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

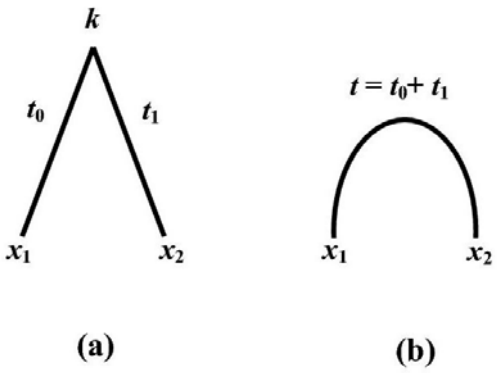


Fig. 5.1. Rooted (a) and unrooted (b) trees for a pair of sequences. Under rev
codon models, the root is unidentifiable; hence, only the sum of the branch l
$t = t_0 + t_1$, is estimable.

Since the ancestral codon is unknown, the summation is over all 61 p
codons for k. Furthermore, as the substitution model is time-reversib
root of the tree can be moved around, say, to species 1, without ch
the likelihood. Thus t_0 and t_1 cannot be estimated individually, an
$t_0 + t_1 = t$ is estimated (Figure 5.1(b)).

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

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

by maximizing the log- likelihood function. Since an analytical solution possible, numerical optimization algorithms are used.

5.2.3 Empirical Demonstration: Pairwise Estimation of the *d* Ratio for *GstD1*

In this section, we use a simple data set and the `codeml` program to trate maximum likelihood estimation of ω. The data set is *GstD1* ge *Drosophila melanogaster* and *D. simulans*. The alignment has 600 c Our first objective is to evaluate the likelihood function for a variety o values for the parameter ω. `Codeml` uses a hill-climbing algorithm to mize the log-likelihood function. In this case, we will let `codeml` estim (`fix_kappa = 0` in the control file `codeml.ctl`) and the sequence di t, but with parameter ω fixed (`fix_omega = 1`). All that remains is `codeml` several times, each with a different value for `omega` in the contr the data in Figure 5.2 show the results for ten different values of ω that the maximum likelihood value for ω appears to be roughly 0.06, w consistent with purifying selection, and that values greater than 1 have lower likelihood scores.

Our second objective is to allow `codeml` to use the hill-climbing algo to maximize the log-likelihood function with respect to κ, t, and ω. Th use `fix_omega = 1` and can use any positive value for `omega`, which i only as a starting value for the iteration. Such a run gives the estimat of 0.067.

Alternatives to maximum likelihood estimates of ω are common [39]. Those methods count the number of sites and differences and then a multiple-hit correction, and they are termed the counting methods. M them make simplistic assumptions about the evolutionary process and ad hoc treatments to the data that can't be justified [23, 39]. Here the *GstD1* sequences to explore the effects of (i) ignoring the transit transversion rate ratio (`fix_kappa = 1`; `kappa = 1`); (ii) ignoring cod age bias (`CodonFreq = 0`); and (iii) alternative treatments of unequal frequencies (`CodonFreq = 2` and `CodonFreq = 3`). Note that for thes transitions are occurring at higher rates than transversions, and cod quencies are very biased, with average base frequencies of 6% (T), 50 5% (A), and 39% (G) at the third position of the codon. Thus, we estimates that account for both biases will be the most reliable.

Results of our exploratory analyses (Table 5.2.3) indicate that mo sumptions are very important for these data. For example, ignoring the tion to transversion ratio almost always led to underestimation of the n of synonymous sites (S), overestimation of d_S, and underestimation of ω is because transitions at the third codon positions are more likely to b onymous than are transversions [19]. Similarly, biased codon usage i

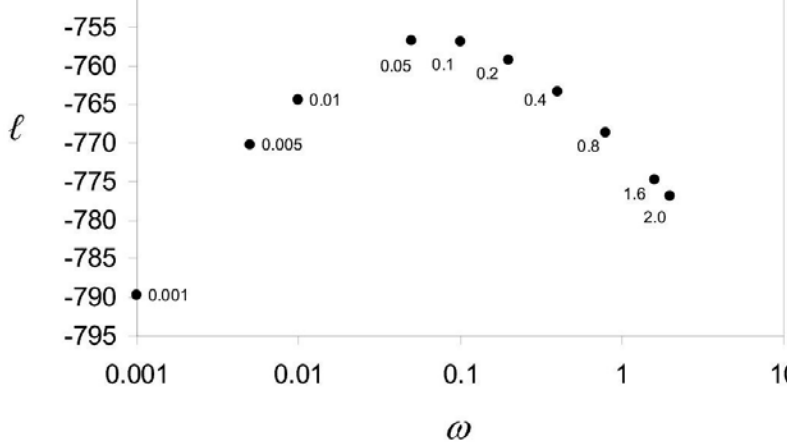


Fig. 5.2. Log-likelihood as a function of the ω parameter for a pair of *GstD* from *Drosophila melanogaster* and *D. simulans*. The maximum likelihood es of ω is the value that maximizes the likelihood function. Since an analytical s is not possible, the `codeml` program uses a numerical hill-climbing algori maximize l. For these data, the maximum likelihood estimate of ω is 0.067, maximum likelihood of -756.57.

unequal substitution rates between the codons, and ignoring it also le biased estimates of synonymous and nonsynonymous substitution ra real data analysis, codon usage bias was noted to have an even greater i than the transition/transversion rate ratio and is opposite to that of ig transition bias. This is clearly indicated by the sensitivity of S to codo where S in this gene (45.2) is less than one-third the expected value the assumption of no codon bias ($S = 165.8$). The estimates of ω diffe much as 4.7-fold (Table 5.2.3). Note that these two sequences differed 3% of sites.

For comparison, we included estimates obtained from two counting ods. The method of Nei and Gojobori [25] is similar to ML ignoring tra bias and codon bias, whereas the method of Yang and Nielsen [39] is sim ML accommodating transition bias and codon bias (F3×4). Note that e tion according to Nei and Gojobori [25] was accomplished by using the c program and according to Yang and Nielsen [39] by using the YN00 pr of PAML. What is clear from these data is that when sequence diverg not too great, assumptions appear to matter more than methods, wi and the counting methods giving similar results under similar assum This result is consistent with simulation studies examining the performa

Method	κ	S	N	d_S	d_N	ω	ℓ
ML methods							
Fequal, $\kappa = 1$	1	152.9	447.1	0.0776	0.0213	0.274	-927.18
Fequal, κ estimated	1.88	165.8	434.2	0.0221	0.0691	0.320	-926.28
F3×4, $\kappa = 1$	1	70.6	529.4	0.1605	0.0189	0.118	-844.51
F3×4, κ 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, κ 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 (F3×4)	3.28	76.6	523.5	0.1499	0.0190	0.127	

different estimation methods [39]. However, as sequence divergence inc
ad hoc treatment of the data can lead to serious estimation errors [23,

5.3 Phylogenetic Estimation of Selective Pressure

Adaptive evolution is very difficult to detect using the pairwise appro
estimating the d_N/d_S ratio. For example, a large-scale database survey
fied less than 1% of genes (17 out of 3595) as evolving under positive se
pressure [9]. The problem with the pairwise approach is that it averages
tive pressure over the entire evolutionary history separating the two li
and over all codon sites in the sequences. In most functional genes, the
ity of amino acid sites will be subject to strong purifying selection [31, 6
only a small fraction of the sites potentially targeted by adaptive evo
[11]. In such cases, averaging the d_N/d_S ratio over all sites will yield
much less than one, even under strong positive selective pressure at
sites. Moreover, if a gene evolved under purifying selection for most
time, with only brief episodes of adaptive evolution, averaging over t
tory of two distantly related sequences would be unlikely to produce a
ratio greater than one [4]. Clearly, the pairwise approach has low po
detect positive selection. Power is improved if selective pressure is allo
vary over sites or branches [37, 40]. However, increasing the complexity
codon model in this way requires that likelihood be calculated for m
sequences on a phylogeny.

Likelihood calculation on a phylogeny (Figure 5.3) is an extension
calculation for two lineages. As in the case of two sequences, the root
be identified and is fixed at one of the ancestral nodes arbitrarily. For ex
given an unrooted tree with four species and two ancestral codons, k
the probability of observing the data at codon site h, $x_h = \{x_1, x_2,$
(Figure 5.3), is

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

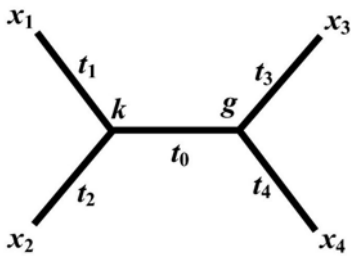


Fig. 5.3. An unrooted phylogeny for four sequences. As in the case of two seq
the root cannot be identified. For the purpose of likelihood calculation, the
fixed at one of the ancestral nodes arbitrarily, and t_0, t_1, t_2, t_3, and t_4 are est
parameters in the model.

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

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

As in the two-species case, numerical optimization is used to ma
the likelihood function with respect to κ, ω, and the $(2N-3)$ branch-
parameters (t's).

5.3.2 Modelling Variable Selective Pressure among Lineages

Adaptive evolution is most likely to occur in an episodic fashion. For
ple, functional divergence of duplicated genes [43, 29, 5], colonizatio

improve detection of episodic adaptive evolution, Yang [37] (see also [2
plemented models that allow for different ω parameters in different par
phylogeny. The simplest model, described above, assumes the same ω ra
all branches in the phylogeny. The most general model, called the "free
model," specifies an independent ω ratio for each branch in a phyloge
the `codeml` program, users can specify an intermediate model, with in
dent ω parameters for different sets of branches. Modelling variable se
pressure involves a straightforward modification of the likelihood cor
tion [37]. Consider the example tree of fig. 5.4. Suppose we suspect se
pressure has changed in one part of this tree, perhaps due to positive se
pressure. To model this, we specify independent ω ratios (ω_0 and ω_1)
two different sets of branches (Figure 5.4). The transition probabilities
two sets of branches are calculated from different rate matrices (Q) gen
by using different ω ratios. Under this model (Figure 5.4), the probab
observing the data at codon site x_h is

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

The log-likelihood function remains a sum over all sites but is now
imized with respect to ω_0 and ω_1, as well as branch lengths (t's) an
parameters for user-defined sets of branches are specified by `model`
the control file and by labelling branches in the tree, as described
PAML documentation.

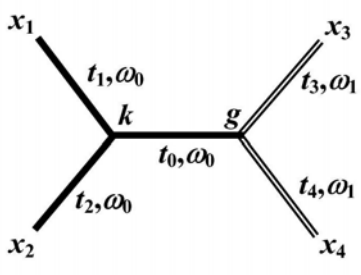


Fig. 5.4. Four-taxon phylogeny with variable ω ratios among its branche
likelihood of this tree is calculated according to Yang [37], where the two in
dent ω ratios (ω_0 and ω_1) are used to calculate rate matrices (Q) and tra
probabilities for the different branches.

In practice, modelling variable selective pressure among sites appears
vide much greater gains in power than does modelling variable selectiv
sure among branches [38]. This is because adaptive evolution is ge
restricted to a small subset of sites [6, 40], and the previous model for
tion over branches effectively averages over all sites. Although differe
the relative rate of nonsynonymous substitution often can be detected
branches, averaging over sites means it is unlikely that estimated ω's
greater than one. In fact, implementation of models with variable ω's
codon sites [26, 40, 41] has led to the detection of positive selection in
genes for which it had not previously been observed. For example, Zan
al. [42] used the models of Nielsen and Yang [26] to detect positive selec
the *nef* gene of HIV-1, a gene for which earlier studies had found no ev
for adaptive evolution [28, 7].

There are two approaches to modelling variation in ω among sites:
a statistical distribution to model the random variation in ω over site
(ii) use a priori knowledge of a protein's structural and functional dom
partition sites in the protein and use different ω's for different partitions
structural and functional information are unknown for most proteins,
tistical distribution will be the most common approach. Collectively, N
and Yang [26] and Yang et al. [40] implemented 13 such models, avail
the `codeml` program. The continuous distributions are approximated by
discrete categories. In this approach, codon sites are assumed to fall i
classes, with the ω ratios for the site classes, and their proportions (p
mated from the data. The number of classes (K) is fixed beforehand, a
ω's and p's are either treated as parameters or functions of parameters
ω distribution [40]. We illustrate likelihood calculation by taking the d
model (M3) as an example. M3 classifies codon sites into K discrete
($i = 0, 1, 2, \ldots, K-1$), with d_N/d_S ratios and proportions given as:

$$\omega_0, \omega_1, ..., \omega_{K-1},$$
$$p_0, p_1, ..., p_{K-1}.$$

Equation (5.4) is used to compute the conditional probability $f(x_h$
the data at a site, h, for each site class. Since we do not know to whic
site h belongs, we sum over both classes, giving the unconditional prob

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

In this way, the unconditional probability is an average over the site
of the ω distribution. Still, assuming that the substitution process at in
ual codon sites is independent, the log-likelihood function is a sum o
sites in the sequence:

The log-likelihood is now maximized as a function of the parame
the ω distribution, branch-lengths (t), and κ.

With the second approach, we used knowledge of a protein's stru
or functional domains to classify codon sites into different partition
different ω's. Since we assume site independence, the likelihood calcula
straightforward; the transition probabilities in equation (5.4) are con
by using the appropriate ω parameter for each codon site. By takin
approach, we are effectively assuming our knowledge of the protein is w
error; hence, we do not average over site classes for each site [41].

5.4 Detecting Adaptive Evolution in Real Data Sets

Maximum likelihood estimation of selective pressure is only one part
problem of detecting adaptive evolution in real data sets. We also ne
tools to rigorously test hypotheses about the nature of selective pressu
example, we might want to test whether d_N is higher than d_S (*i.e.*, ω
Fortunately, we can combine estimation of selective pressure with a
statistical approach to hypothesis testing, the likelihood ratio test (
Combined with Markov models of codon evolution, the LRT provides
general method for testing hypotheses about protein evolution, includin
test for variation in selective pressure among branches; (ii) a test for va
in selective pressure among sites; and (iii) a test for a fraction of sites ev
under positive selective pressure. In the case of a significant LRT fo
evolving under positive selection, we use Bayes or empirical Bayes m
to identify positively selected sites in an alignment. In the following s
we provide an introduction to the LRT and Bayes' theorem and provid
empirical demonstrations of their use on real data.

5.4.1 Likelihood Ratio Test (LRT)

The LRT is a general method for testing assumptions (model paran
through comparison of two competing hypotheses. For our purposes, w
only consider comparisons of nested models; that is, where the null hyp
(H_0) is a restricted version (special case) of the alternative hypothesis
Note that the LRT only evaluates the differences between a pair of m
and any inadequacies shared by both models remain untested. Let ℓ_0
maximum log-likelihood under H_0 with parameters θ_0, and let ℓ_1 be th
imum log-likelihood under H_1 with parameters θ_1. The log-likelihood st
is defined as twice the log likelihood difference between the two model

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

between the two models.

Use of the χ^2 approximation to the likelihood ratio statistic require certain conditions be met. First, the hypotheses must be nested. Secon sample must be sufficiently large; the χ^2 approximation fails when too fe are used. Third, H_1 may not be related to H_0 by fixing one or more parameters at the boundary of parameter space. This is called the "boun problem, and the LRT statistic is not expected to follow a χ^2 distribu this case [30]. When the conditions above are not met, the exact distri can be obtained by Monte Carlo simulation [12, 1], although this ca computationally costly solution.

5.4.2 Empirical Demonstration: LRT for Variation in Selectiv Pressure among Branches in *Ldh*

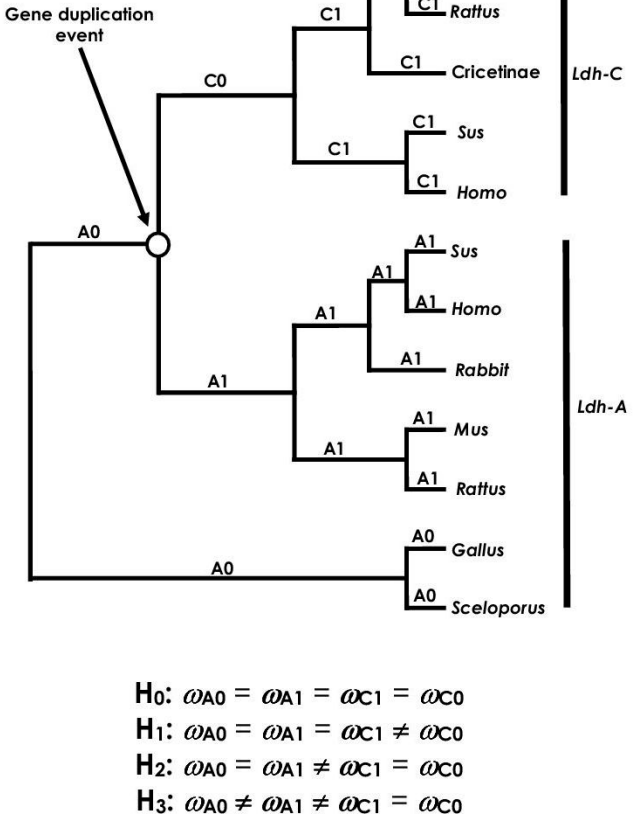
The *Ldh* gene family is an important model system for molecular evo of isozyme multigene families [20]. The paralogous copies of lactate de genase (*Ldh*) genes found in mammals originated from a duplication ne origin of vertebrates (*Ldh-A* and *Ldh-B*) and a later duplication near t gin of mammals (Figure 5.5; *Ldh-A* and *Ldh-C*). Li and Tsoi [20] foun the rate of evolution had increased in mammalian *Ldh-C* sometime fol the second duplication event. An unresolved question about this gene fa whether the increased rate of *Ldh-C* reflects (i) a burst of positive select functional divergence following the duplication event, (ii) a long-term c in selective pressure, or (iii) simply an increase in the underlying mu rate of *Ldh-C*. In the following, we use the LRT for variable ω ratios branches to test these evolutionary scenarios.

The null hypothesis (H_0) is that the rate increase in *Ldh-C* is due to an underlying increase in the mutation rate. If the selective pr was constant and the mutation rate increased, the relative fixation ra synonymous and nonsynonymous mutations (ω) would remain constan the phylogeny, but the overall rate of evolution would increase in *Ldh-C* alternative to this scenario is that the rate increase in *Ldh-C* was du burst of positive selection following gene duplication (H_1). A formal t variation in selective pressure among sites may be formulated as follow

H_0: ω is identical across all branches of the *Ldh* phylogeny.

H_1: ω is variable, being greater than 1 in branch $C0$ of Figure 5.5.

Because H_1 can be transformed into H_0 by restricting ω_{C0} to be to the ω ratios for the other branches, we can use the LRT. The estim ω under the null hypothesis, as an average over the phylogeny in Figu was 0.14, indicating that evolution of *Ldh-A* and *Ldh-C* was domina purifying selection. The LRT suggests that selective pressure in *Ldh-C* diately following gene duplication (0.19) was not significantly differen the average over the other branches (Table 5.2). Hence, we found no ev



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

Fig. 5.5. A phylogenetic tree for the *Ldh-A* and *Ldh-C* gene families. T was obtained by a neighbor-joining analysis of a codon sequence alignment the HKY85 substitution model [14] combined with a Gamma model of ra ation among sites [35]. Branch lengths are not to scale. The *Gallus* (chicke *Sceloporus* (lizard) *Ldh-A* sequences are pro-orthologs, as they predate th duplication event. The tree is rooted with the pro-orthologous sequences for nience; all analyses were conducted by using the unrooted topology. The or model (H_0) assumes uniform selective pressure over all branches. H_1 is based notion of a burst of positive selection in *Ldh-C* following the gene duplicatio hence the assumption of one ω for branch $C0$ and another for all other br H_2 is based on the notion of increased nonsynonymous substitution in all lineages following gene duplication; hence the assumption of one ω for the branches ($\omega_{C0} = \omega_{C1}$) and another for the *Ldh-A* branches ($\omega_{A0} = \omega_{A1}$). H_3 i on the notion that selective pressure changed in both *Ldh-C* and *Ldh-A* fo gene duplication, as compared with the pro-orthologous sequences; hence, or the *Ldh-C* branches ($\omega_{C0} = \omega_{C1}$), one ω for the post-duplication *Ldh-A* br (ω_{A1}), and one ω for the pro-orthologous branches (ω_{A0}).

selection for just one or a few amino acid changes, we would not obs
large difference in ω ratios among branches.

Table 5.2. Parameter estimates under models of variable ω ratios among l
for the *Ldh-A* and *Ldh-C* gene families. (Note: The topology and branch-spe
ratios are presented in Figure 5.5. The df is 1 for the comparisons of H_0 vs.
vs. H_2, and H_2 vs. H_3.)

Models	w_{A0}	w_{A1}	w_{C1}	w_{C0}	ℓ
$H_0 : w_{A0} = w_{A1} = w_{C1} = w_{C0}$	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

Using the same approach, we tested a second alternative hypothesis,
the rate increase in *Ldh-C* was due to an increase in the nonsynon
substitution rate over all lineages of the *Ldh-C* clade (see H_2 in Figu
In this case, the LRT was highly significant, and the parameter estima
the *Ldh-C* clade indicated an increase in the relative rate of nonsynon
substitution by a factor of 3 (Table 5.2). Lastly, we tested the hypothes
selective pressure differed in both *Ldh-A* and *Ldh-C* following gene dupl
(see H_3 in Figure 5.5), and results of this test were not significant
5.2). Collectively, these findings suggest selective pressure and mutatio
in *Ldh-A* were relatively unchanged by the duplication event, where
nonsynonymous rate increased in *Ldh-C* following the duplication ev
compared with *Ldh-A*.

5.4.3 Empirical Demonstration: Positive Selection in the *nef* in the Human HIV-2 Genome

The role of the *nef* gene in differing phenotypes of HIV-1 infection ha
well-studied, including identification of sites evolving under positive se
pressure [42]. The *nef* gene in HIV-2 has received less attention, presu
because HIV-2 is associated with reduced virulence and pathogenicity r
to HIV-1. Padua et al. [27] sequenced 44 *nef* alleles from a study popu
of 37 HIV-2-infected people living in Lisbon, Portugal. They found th
cleotide variation in the *nef* gene, rather than gross structural chang
potentially correlated with HIV-2 pathogenesis. In order to determine w
the *nef* gene might also be evolving under positive selective pressure i
2, we analyzed those same data here with models of variable ω ratios
sites [40].

M3 (discrete), M7 (beta), and M8 (beta & ω). Models M0 and M
described above. M1 (neutral) specifies two classes of sites: conserve
with $\omega = 0$ and neutral sites with $\omega = 1$. M2 (selection) is an extension
(neutral), adding a third ω class that is free to take a value > 1. Versio
of `paml/codeml` introduces a slight variation to models M1 (neutral) a
(selection) in that $\omega_0 < 1$ is estimated from the data rather than bein
at 0. Those are referred to as models M1a and M2a, also used here.
model M7 (beta), ω varies among sites according to a beta distributio
parameters p and q. The beta distribution is restricted to the interval
thus, M1 (neutral), M1a (nearly neutral), and M7 (beta) assume no p
selection. M8 (beta & ω) adds a discrete ω class to the beta distributio
is free to take a value > 1. Under M8 (beta & ω), a proportion of s
is drawn from a beta distribution, with the remainder ($p_1 = 1 - p_0$)
the ω ratio of the added site class. We specified $K = 3$ discrete cla
sites under M3 (discrete), and $K = 10$ under M7 (beta) and M8 (b
ω). We use an LRT comparing M0 (one ratio) with M3 (discrete) to t
variable selective pressure among sites and three LRTs to test for sites ev
by positive selection, comparing (i) M1 (neutral) against M2 (selectio
M1a (nearly neutral) and M2a (positive selection), and (iii) M7 (beta) a
M8 (beta & ω).

Maximum likelihood estimates of parameters and likelihood scores
nef gene are presented in Table 5.3. Averaging selective pressure ove
and branches as in M0 (one ratio) yielded an estimated ω of 0.50, a
consistent with purifying selection. The LRT comparing M0 (one ratio) a
M3 (discrete) is highly significant ($2\Delta\ell = 1087.2$, df $= 4$, $P < 0.01$), indi
that the selective pressure is highly variable among sites. Estimates of ω
models that can allow for sites under positive selection (M2, M2a, M
indicated a fraction of sites evolving under positive selective pressure
5.3). To formally test for the presence of sites evolving by positive sel
we conducted LRTs comparing M1 and M2, M1a and M2a, and M
M8. All those LRTs were highly significant; for example, the test statis
comparing M1 (neutral) and M2 (selection) is $2\Delta\ell = 223.58$, with $P <$
df $= 2$. These findings suggest that about 12% of sites in the *nef* g
HIV-2 are evolving under positive selective pressure, with ω between
3. It is clear from Table 5.3 that this mode of evolution would not hav
detected if ω were measured simply as an average over all sites of *nef*.

Models M2 (selection) and M8 (beta & ω) are known being multipl
optima in some data sets, often with ω_2 under M2 or ω under M8 to be
one peak and > 1 on another peak. Thus it is important to run these
multiple times with different starting values (especially different ω's) an
select the set of estimates corresponding to the highest peak. Indeed, t
dataset illustrates this issue. By using different initial ω's, both the glob
local optima can be found.

in parentheses), is the number of free parameters in the ω distribution. The
ratio is an average over all sites in the HIV-2 *nef* gene alignment. Parame
parentheses are not free parameters and are presented for clarity. PSS is the n
of positive selected sites, inferred at the 50% (95%) posterior probability cu

Model	d_N/d_S	Parameter estimates	PSS	
M0: one ratio (1)	0.51	$\omega = 0.505$	none	−9
M3: discrete (5)	0.63	$p_0 = 0.48, p_1 = 0.39, (p_2 = 0.13)$ $\omega_0 = 0.03, \omega_1 = 0.74, \omega_2 = 2.50$	31 (24)	−9
M1: neutral (1)	0.63	$p_0 = 0.37, (p_1 = 0.63)$ $(\omega_0 = 0), (\omega_1 = 1)$	not allowed	−9
M2: selection (3)	0.93	$p_0 = 0.37, p_1 = 0.51, (p_2 = 0.12)$ $(\omega_0 = 0), (\omega_1 = 1), \omega_2 = 3.48$	30 (22)	−9
M1a: nearly neutral (2)	0.48	$p_0 = 0.55, (p_1 = 0.45)$ $(\omega_0 = 0.06), (\omega_1 = 1)$	not allowed	−9
M2a: positive selection (4)	0.73	$p_0 = 0.51, p_1 = 0.38, (p_2 = 0.11)$ $(\omega_0 = 0.05), (\omega_1 = 1), \omega_2 = 3.00$	26 (15)	−9
M7: beta (2)	0.42	$p = 0.18, q = 0.25$	not allowed	−9
M8: beta & ω (4)	0.62	$p_0 = 0.89, (p_1 = 0.11)$ $p = 0.20, q = 0.33, \omega = 2.62$	27 (15)	−9

5.4.4 Bayesian Identification of Sites Evolving under Positive Darwinian Selection

Under the approach described in this chapter, a gene is considered t
evolved under positive selective pressure if (i) the LRT is significant a
at least one of the ML estimates of $\omega > 1$. Given that these conditio
satisfied, we have evidence for sites under positive selection but no in
tion about which sites they are. Hence, the empirical Bayes approach
to predict them [26, 40]. To do this, we compute, in turn, the posterior
ability of a site under each ω site class of a model. Sites with high po
probabilities under the class with $\omega > 1$ are considered likely to have e
under positive selective pressure.

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

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

Because the parameters used in the equation above to calculate th
terior probability are estimated by ML (ω_i and p_i), the approach is
empirical Bayes. By using the ML parameters in this way, we ignor

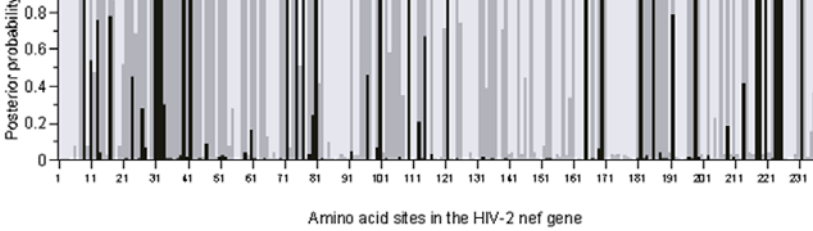


Fig. 5.6. Posterior probabilities for sites classes under M3 ($K = 3$) along the *nef* gene alignment.

sampling errors. The posterior probabilities will be sensitive to these pa ter estimates, meaning that the reliability of this approach will be poor the parameter estimates are poor, such as in small datasets or when ob from a local optimum.

Because the *nef* dataset above is quite large, the parameter est are expected to be reliable [2]. Consistent with this, ML estimates strength and proportion of positively selected sites in *nef* are consistent M2, M3, and M8 (Table 5.3). Figure 5.6 shows the posterior probabilit the $K = 3$ site classes at each site of *nef* under model M3. Twent sites were identified as having very high posterior probability ($P > 0$ evolving under positive selection (site class with $\omega > 1$). Interestingly of these sites matched the two variable sites in a proline-rich motif strongly associated with an asymptomatic disease profile [27]. In fact four of the 24 sites were found in regions of *nef* considered importa function. Disruption of the important *nef* regions is associated with re pathogenicity in HIV-2-infected individuals [32, 27]. Our results sugges selective pressure at such sites is fundamentally different from selection at the 24 positive selection sites predicted using the Bayes theorem. identified with such high posterior probabilities, the predicted sites mus been evolving under long-term positive selective pressure, suggesting tha are more likely subjected to immune-driven diversifying selection at ep [42, 34].

5.5 Power, Accuracy and Robustness

The boundary problem mentioned above applies to the LRT for varia lective pressure among sites and the LRT for positive selection at a frac sites [1]. The problem arises in the former because the null (M0) is equ to M3 ($K = 3$) with two of the five parameters (p_0 and p_1) fixed to 0,

proportion parameter (p) fixed to 0. Therefore, the χ^2 approximation
expected to hold. Anisimova et al. [1] used computer simulation to inve
the effect of the boundary problem on the power and accuracy of the
Use of the χ^2 makes the LRT conservative, meaning that the false p
rate will be less than predicted by the specified significance level of t
[1]. Nevertheless, the test was found to be powerful, sometimes reaching
in data sets consisting of 17 sequences. Power was low for highly simil
highly divergent sequences but was modulated by the length of the se
and the strength of positive selection. Note that simulation studies, bot
and without the boundary problem, indicate that the sample size re
ments for the χ^2 approximation are met with relatively short sequen
some cases as few as 50 codons [1].

Bayesian prediction of sites evolving under positive selection is a
difficult task than ML parameter estimation or likelihood ratio testin
difficulty arises because the posterior probabilities depend on the (i) in
tion contained at just a single site in the data set and (ii) the quality of t
parameter estimates. Hence, a second study was conducted by Anisim
al. [2] to examine the power and accuracy of the Bayesian site identifi
The authors made the following generalizations: (i) prediction of pos
selected sites is not practical from just a few highly similar sequenc
the most effective method of improving accuracy is to increase the num
lineages; and (iii) site prediction is sensitive to sampling errors in para
estimates and to the assumed ω distribution.

Robustness refers to the stability of results to changes in the mo
sumptions. The LRT for positive selection is generally robust to the as
distribution of ω over sites [1]. However, as the LRT of M0 with M3 is a
variable selective pressure among sites, caution must be exercised whe
the M0–M3 comparison suggests positive selection. One possibility is
M2, which tends to be more conservative than the other models [2]. A
approach is to select the subset of sites that are robust to the ω distri
[1, 34]. A third approach is to select sites that are robust to sampling li
[34]. We believe that sensitivity analysis is a very important part of de
positive selection, and we make the following recommendations: (i) m
models should be used, (ii) care should be taken to identify and discard
obtained from local optima, and (iii) assumptions such as the ω distri
or the method of correcting for biased codon frequencies should be
ated relative to their effects on ML parameter estimation and Bayesia
prediction.

All codon models discussed above ignore the effect of the physicoch
property of the amino acid being substituted. For example, all amin
substitutions at a positively selected site are assumed to be advanta
with $\omega > 1$. The assumption appears to be unrealistic; one can imagin
there might be a set of amino acid substitutions that are forbidden at

conservative, only indicating positive selection when the estimate of ω
In cases where only one or a few amino acid substitutions result in a su
tial change in phenotype, the methods will have little or no power bec
will be < 1. Another important limitation is the assumption of a single
lying phylogeny. When recombination has occurred, no single phyloge
fit all sites of the data. A recent simulation study [3] found that the
robust to low levels of recombination but can have a seriously high typ
ror rate when recombination is frequent. Interestingly, Bayesian predic
positively selected sites was less affected by recombination than was the
In summary, no matter how robust the results, they must be interprete
these limitations in mind.

Acknowledgment

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

References

[1] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and power
likelihood ratio test in detecting adaptive molecular evolution. *Mo
Evol.*, 18:1585–1592, 2001.

[2] M. Anisimova, J. P. Bielawski, and Z. Yang. Accuracy and po
bayesian prediction of amino acid sites under positive selection.
Biol. Evol., 19:950–958, 2002.

[3] M. Anisimova, R. Nielsen, and Z. Yang. Effect of recombination
accuracy of likelihood methods for detecting positive selection at
acid sites. *Genetics*, 164:1229–1236, 2003.

[4] J. P. Bielawski and Z. Yang. The role of selection in the evolution
daz gene family. *Mol. Biol. Evol.*, 18:523–529, 2001.

[5] J. P. Bielawski and Z. Yang. Maximum likelihood methods for de
adaptive evolution after gene duplication. *J. Struct. Funct. Gen
3:201–212, 2003.

[6] K. A. Crandall, C. R. Kelsey, H. Imanichi, H. C. Lane, and N. P
man. Parallel evolution of drug resistance in HIV: Failure of nons
mous/synonymous substitution rate ratio to detect selection. *Mo
Evol.*, 16:372–382, 1999.

[7] J. Da Silva and A. L. Hughes. Conservation of cytotoxic t lymph
(CTL) epitopes as a host strategy to constrain parasitic adaptatio
idence from the *nef* gene of human immunodeficiency virus 1 (H
Mol. Biol. Evol., 15:1259–1268, 1998.

Genetics, 157:295–305, 2001.

[9] T. K. Endo, K. Ikeo, and T. Gojobori. Large-scale search for ge
which positive selection may operate. *Mol. Biol. Evol.*, 13:685–690

[10] J. Felsenstein. Evolutionary trees from DNA sequences: A ma
likelihood approach. *J. Mol. Evol.*, 15:368–376, 1981.

[11] G. B. Golding and A. M. Dean. The structural basis of molecular
tation. *Mol. Biol. Evol.*, 15:355–369, 1998.

[12] N. Goldman. Statistical tests of DNA substitution models. *J. Mol.*
36:182–198, 1993.

[13] N. Goldman and Z. Yang. A codon based model of nucleotide subst
for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736,

[14] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape sp
by a molecular clock using mitochondrial DNA. *J. Mol. Evol.*, 2
174, 1985.

[15] Y. Ina. Pattern of synonymous and nonsynonymous substitutio
indicator of mechanisms of molecular evolution. *J. Genet.*, 75:9
1996.

[16] F. M. Jiggins, G. D. D. Hurst, and Z. Yang. Host-symbiont co
Positive selection on the outer membrane protein of parasite b
mutualistic *Rickettsiaceae*. *Mol. Biol. Evol.*, 19:1341–1349, 2002.

[17] M. Kimura and T. Ohta. On some principles governing molecular
tion. *Proc. Natl. Acad. Sci. USA*, 71:2848–2852, 1974.

[18] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inf
of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.*, 3
160, 1990.

[19] W.-H. Li, C.-I. Wu, and C.-C. Luo. A new method for estimating s
mous and nonsynonymous rates of nucleotide substitutions consi
the relative likelihood of nucleotide and codon changes. *Mol. Biol.*
2:150–174, 1985.

[20] Y.-J. Li and C.-M. Tsoi. Phylogenetic analysis of vertebrate lacta
hydrogenase (ldh) multigene families. *J. Mol. Evol.*, 54:614–624, 2

[21] W. Messier and C.-B. Stewart. Episodic adaptive evolution of p
lysozymes. *Nature*, 385:151–154, 1997.

[22] T. Miyata and T. Yasunaga. Molecular evolution of mRNA: A m
for estimating evolutionary rates of synonymous and amino acid
tutions from homologous nucleotide sequences and its applicatio
Mol. Evol., 16:23–36, 1980.

[23] S. V. Muse. Estimating synonymous and non-synonymous substi
rates. *Mol. Biol. Evol.*, 13:105–114, 1996.

[24] S. V. Muse and B. S. Gaut. A likelihood approach for comparin
onymous and nonsynonymous nucleotide substitution rates, with
cations to the chloroplast genome. *Mol. Biol. Evol.*, 11:715–725, 1

Evol., 3:418–426, 1986.

[26] R. Nielsen and Z. Yang. Likelihood models for detecting positiv lected amino acid sites and applications to the HIV-1 envelope *Genetics*, 148:929–936, 1998.

[27] E. Padua, A. Jenkins, S. Brown, J. Bootman, M. T. Paixao, N. Al and N. Berry. Natural variation of the *nef* gene in human immu ciency virus type 2 infections in Portugal. *J. Gen. Virol.*, 84:1287 2003.

[28] U. Plikat, K. Nieselt-Struwe, and A. Meyerhans. Genetic drift can mine short-term human immunodeficiency virus type 1 *nef* quasis evolution in vivo. *J. Virol.*, 71:4233–4240, 1997.

[29] T. R. Schmidt, M. Goodman, and L. I. Grossman. Molecular evo of the cox7a gene family in primates. *Mol. Biol. Evol.*, 16:619–626

[30] S. Self and K. Y. Liang. Asymptotic properties of maximum like estimators and likelihood ratio tests under non-standard conditio *Am. Stat. Assoc.*, 82:605–610, 1987.

[31] P. M. Sharp. In search of molecular Darwinism. *Nature*, 385:11 1997.

[32] W. M. Switzer and S. Wiktor et al. Evidence of *nef* truncation in l immunodeficiency virus type 2 infection. *J. Infect. Dis.*, 177:65–71

[33] M. Wayne and K. Simonsen. Statistical tests of neutrality in an weak selection. *Trends Ecol. Evol.*, 13:236–240, 1998.

[34] W. Yang, J. P. Bielawski, and Z. Yang. Widespread adaptive evo in the human immunodeficiency virus type 1 genome. *J. Mol.* 57(2):212–221, 2003.

[35] Z. Yang. Maximum likelihood phylogenetic estimation from DN quences with variable rates over sites: Approximate methods. *J Evol.*, 39:306–314, 1994.

[36] Z. Yang. PAML: A program package for phylogenetic analysis by mum likelihood. *Appl. Biosci.*, 13:555–556, 1997.

[37] Z. Yang. Likelihood ratio tests for detecting positive selection a plication to primate lysozyme evolution. *Mol. Biol. Evol.*, 15:56 1998.

[38] Z. Yang and J. P. Bielawski. Statistical methods for detecting mo adaptation. *Trends Ecol. Evol.*, 15:496–503, 2000.

[39] Z. Yang and R. Nielsen. Estimating synonymous and nonsynon substitution rates under realistic evolutionary models. *Mol. Biol.* 17:32–43, 2000.

[40] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen. C substitution models for heterogeneous selective pressure at amin sites. *Genetics*, 155:431–449, 2000.

site classes. *Mol. Biol. Evol.*, 19:49–57, 2002.

[42] P. M. Zanotto, E. G. Kallis, R. F. Souza, and E. C. Holmes. Ge
ical evidence for positive selection in the *nef* gene of HIV-1. *Ge*
153:1077–1089, 1999.

[43] J. Zhang, H. F. Rosenberg, and M. Nei. Positive Darwinian se
after gene duplication in primate ribonuclease genes. *Proc. Natl.*
Sci. USA, 95:3708–3713, 1998.