Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection

Ziheng Yang, Wendy S.W. Wong,† and Rasmus Nielsen†‡*

*Department of Biology, University College London, London, United Kingdom; †Department of Biological Statistics and Computational Biology, Cornell University; and ‡Center for Bioinformatics, University of Copenhagen, Copenhagen, Denmark

Codon-based substitution models have been widely used to identify amino acid sites under positive selection in comparative analysis of protein-coding DNA sequences. The nonsynonymous-synonymous substitution rate ratio (d_N/d_S, denoted ω) is used as a measure of selective pressure at the protein level, with $\omega > 1$ indicating positive selection. Statistical distributions are used to model the variation in ω among sites, allowing a subset of sites to have $\omega > 1$ while the rest of the sequence may be under purifying selection with $\omega < 1$. An empirical Bayes (EB) approach is then used to calculate posterior probabilities that a site comes from the site class with $\omega > 1$. Current implementations, however, use the naive EB (NEB) approach and fail to account for sampling errors in maximum likelihood estimates of model parameters, such as the proportions and ω ratios for the site classes. In small data sets lacking information, this approach may lead to unreliable posterior probability calculations. In this paper, we develop a Bayes empirical Bayes (BEB) approach to the problem, which assigns a prior to the model parameters and integrates over their uncertainties. We compare the new and old methods on real and simulated data sets. The results suggest that in small data sets the new BEB method does not generate false positives as did the old NEB approach, while in large data sets it retains the good power of the NEB approach for inferring positively selected sites.

Introduction

The nonsynonymous-synonymous substitution rate ratio (d_N/d_S or ω) provides a measure of selective pressure at the protein level (Miyata, Miyazawa, and Yasunaga 1979; Li, Wu, and Luo 1985). An ω greater than one indicates that nonsynonymous mutations offer fitness advantages and are fixed in the population at a higher rate than synonymous mutations. Positive selection can thus be detected by identifying cases where $\omega > 1$. In a functional protein, many amino acids may be under strong structural and functional constraints and not free to vary. Thus, it is important to account for variation in selective pressure (and thus in the ω ratio) among sites if one hopes to detect positive selection affecting only a few amino acid residues (Nielsen and Yang 1998; Suzuki and Gojobori 1999). A number of such models were implemented by Nielsen and Yang (1998) and Yang et al. (2000) based on the codon-substitution model of Goldman and Yang (1994; see also Muse and Gaut 1994). In the past few years, such site-specific models have been used to detect positive selection in a variety of genes and species (e.g., Zanotto et al. 1999; Bishop, Dean, and Mitchell-Olds 2000; Bielawski and Yang 2001; Ford 2001; Haydon et al. 2001; Swanson et al. 2001; Mondragon-Palomino et al. 2002; Twiddy, Woelk, and Holmes 2002; Takebayashi et al. 2003; Filip and Mundy 2004; Lane et al. 2004; Moury 2004). Computer simulations also confirmed the power of those methods (Anisimova, Bielawski, and Yang 2001, 2002; Wong et al. 2004).

Analysis of both real and simulated data has provided insights into the statistical properties of the models and highlighted the strengths and weaknesses of such codon-based analysis. The site models of Nielsen and Yang (1998) and Yang et al. (2000) use a statistical distribution to describe the random variation in ω among sites. A likelihood ratio test (LRT) is conducted to compare a null model that does not allow $\omega > 1$ in the distribution with an alternative model that does. Several LRTs were implemented and two appeared to have good power and low false-positive rate. The first involves the null model M1a (NearlyNeutral), which assumes two site classes in proportions p_0 and $p_1 = 1 - p_0$ with $0 < \omega_0 < 1$ and $\omega_1 = 1$, and the alternative model M2a (PositiveSelection), which adds a proportion p_2 of sites with $\omega_2 > 1$ estimated from the data. Those are slight modifications of models M1 (neutral) and M2 (selection) implemented in Nielsen and Yang (1998), which had $\omega_0 = 0$ fixed. The old M1 and M2 were found to be unrealistic for many data sets as they failed to account for sites under weak purifying selection with $0 < \omega < 1$ (e.g., Yang et al. 2000). The second LRT compares the null model M7 (beta), which assumes a beta distribution for ω (in the interval $0 < \omega < 1$), and the alternative model M8 (beta&ω), which adds an extra class of sites with positive selection ($\omega_s > 1$). If the LRT is significant, positive selection is inferred. An empirical Bayes (EB) approach is then used to calculate the posterior probability that each site is from a particular site class, and sites with high posterior probabilities coming from the class with $\omega > 1$ (say, with $P > 95\%$) are inferred to be under positive selection. This approach makes it possible to detect positive selection and identify sites under positive selection even if the average ω ratio over all sites is much less than 1.

The EB approach we implemented, known as the naive EB (NEB), uses maximum likelihood estimates (MLEs) of parameters, such as the proportions and ω ratios for the site classes, without accounting for their sampling errors. While this is not a problem in large data sets, where parameters are reliably estimated, in small data sets the MLEs may have large sampling errors, and the NEB calculation of posterior probabilities may be unreliable (Anisimova, Bielawski, and Yang 2002). For example, if the MLEs under M2a are $\hat{p}_0=\hat{p}_1=0, \hat{p}_2=1$, and $\hat{\omega}_2=1.3$, use of such estimates to calculate posterior probabilities will lead to the conclusion that every site in the sequence is under positive selection with $P = 1$. Such extreme estimates

Key words: positive selection, codon-substitution models, Bayes empirical Bayes.

E-mail: rasmus@binf.ku.dk.

Mol. Biol. Evol. 22(4):1107–1118. 2005
doi:10.1093/molbev/msi097
Advance Access publication February 2, 2005

can occur, for example, when the data contain a few almost identical sequences.

One solution to this problem was provided by Huelsenbeck and Dyer (2004). They implemented a full Bayesian method for calculating posterior probabilities using Markov Chain Monte Carlo. By assigning prior probabilities to the nuisance parameters, the method is able to take uncertainty in these parameters into account. While this method may have desirable statistical properties, it is computationally slow and may not be practical for large data sets or for evaluation by simulation. Also, the method was implemented only under models M2 and M3 (discrete) (Yang et al. 2000) instead of the more useful models M2a or M8.

In this paper, we develop a method to accommodate uncertainties in the MLEs of parameters in the ω distribution using numerical integration. We assign a prior for those parameters and average over this prior, a procedure known as Bayes empirical Bayes (BEB) (Deely and Lindley 1981). We expect that the effects of this correction should be negligible in large data sets but may be important in small data sets. Thus, we test the new method using three data sets analyzed previously: a large informative data set of 192 human class I major histocompatibility complex (MHC) alleles, analyzed by Yang and Swanson (2002); a data set of HIV-1 *env* gene V3 region from 13 HIV-1 isolates with a known transmission history, analyzed by Yang et al. (2000) (data set D10); and a data set of 20 HTLV-I *tax* gene sequences, analyzed by Suzuki and Nei (2004). We also conduct computer simulation to examine the performance of the new BEB method in comparison with the old NEB method.

We also implement similar BEB corrections for branch-site model A of Yang and Nielsen (2002) and the clade model C of Bielawski and Yang (2004) (see also Forsberg and Christiansen 2003). Our implementations are described in *Methods*. Simulation studies evaluating the performance of those models will be published elsewhere.

Methods

BEB Calculation of Probabilities of Sites Under Positive Selection Under Site-Specific Models

The likelihood method of Nielsen and Yang (1998) and Yang et al. (2000) assumes that the d_N/d_S ratio $\omega^{(h)}$ for site h varies according to a statistical distribution $f(\omega|\eta)$ with parameters η. As discussed above, two LRTs, comparing M1a with M2a and M7 with M8, respectively, appear to have good performance. Thus, in this paper we focus on the two alternative models in those tests: M2a and M8. M2a assumes three site classes in proportions p_0, p_1, and $p_2 = 1 - p_0 - p_1$ with $0 < \omega_0 < 1$, $\omega_1 = 1$, and $\omega_2 > 1$. Thus, $\eta = (p_0, p_1, \omega_0, \omega_2)$. M8 assumes that a proportion p_0 of sites are conserved with $\omega_0 \sim \text{beta}(p, q)$, while the remaining sites (proportion $p_1 = 1 - p_0$) are under positive selection with $\omega_s > 1$. Thus, $\eta = \{p_0, p, q, \omega_s\}$. In either model, parameters η are estimated from the likelihood function

$$f(X|\eta) = \prod_{h=1}^{n} f(x_h|\eta) = \prod_{h=1}^{n} \int f(x_h|\omega^{(h)} = \omega) f(\omega|\eta) d\omega$$
$$= \prod_{h=1}^{n} \sum_{k} f(x_h|\omega^{(h)} = \omega_k) f(\omega_k|\eta), \qquad (1)$$

where $X = \{x_h\}$ is the data or sequence alignment, x_h is the data at site h, with $h = 1, 2, \ldots, n$. The last equality holds when the distribution of ω is discrete. Under M8, the integral over the beta distribution of ω is approximated using 10 equal-probability categories (Yang et al. 2000). Thus, the sum over k is over 3 site classes under M2a and over 11 site classes under M8. When our interest is $\omega^{(h)}$, we can view $f(\omega^{(h)}|\eta)$ as the prior and η as the parameters of the prior. Nielsen and Yang (1998) calculated the posterior probability $P = \Pr(\omega^{(h)} = \omega_k|x_h, \eta)$ with η replaced by the MLE, $\hat{\eta}$. This naive empirical Bayes (NEB) approach fails to account for sampling errors in $\hat{\eta}$. In this paper we develop a correction to take into account uncertainties in η. Other parameters, such as the branch lengths and the transition/transversion rate ratio, appear much less important to the calculation of the posterior probabilities, and their values are fixed at the MLEs.

Several procedures have been suggested in the statistics literature to correct for the bias in the NEB approach to achieve approximately correct frequentist coverage probabilities (e.g., Morris 1983; Laird and Louis 1987; Carlin and Gelfand 1990). Most of them work only for simplistic examples or otherwise involve complicated approximations. Laird and Louis (1987) proposed a general approach to the problem, using what they called the type III parametric bootstrap. However, the approach used for the present problem would involve extensive computation. Here, we take a hierarchical Bayes approach, also known as BEB (Deely and Lindley 1981). We use a prior $f(\eta)$ for parameters η and integrate over the prior.

Thus, for any site h

$$\Pr(\omega^{(h)} = \omega_k|X) = \frac{1}{f(X)} \int f(X|\omega^{(h)} = \omega_k, \eta) f(\omega_k|\eta) f(\eta) d\eta$$
$$= \frac{1}{f(X)} \int f(x_h|\omega^{(h)} = \omega_k) f(\omega_k|\eta)$$
$$\times \prod_{j \neq h} \left[\sum_{k'} f(x_j|\omega^{(j)} = \omega_{k'}) f(\omega_{k'}|\eta) \right] f(\eta) d\eta, \qquad (2)$$

where

$$f(X) = \int \prod_{j=1}^{n} \left[\sum_{k'} f(x_j|\omega^{(j)} = \omega_{k'}) f(\omega_{k'}|\eta) \right] f(\eta) d\eta, \quad (3)$$

is a normalizing constant. Note that the sum over k' is over the 3 site classes under M2a or over the 11 site classes under M8. In equation (2), the product over j gives the probability of observing data at all sites except site h. In the NEB approach, where parameters η are fixed, data at other sites do not provide information about ω at site h so that $\Pr(\omega^{(h)}|X, \eta) = \Pr(\omega^{(h)}|x_h, \eta)$. However, for the BEB, this is not the case so that we have to consider $\Pr(\omega^{(h)}|X)$.

We approximate the integral over η by a sum over a 4-D grid.

$$\Pr(\omega^{(h)} = \omega_k|X) = \frac{1}{f(X)} \sum_{s} f(x_h|\omega^{(h)} = \omega_k) f(\omega_k|\eta_s)$$
$$\times \prod_{j \neq h} \left[\sum_{k'} f(x_j|\omega^{(j)} = \omega_{k'}) f(\omega_{k'}|\eta_s) \right] f(\eta_s), \qquad (4)$$

where

$$f(X) = \sum_s \left\{ \prod_{j=1}^{n} \left[\sum_{k'} f(x_j|\omega^{(j)} = \omega_{k'}) f(\omega_{k'}|\eta_s) \right] f(\eta_s) \right\}. \tag{5}$$

The posterior mean and variance of $\omega^{(h)}$ can be calculated similarly. For example,

$$E(\omega^{(h)}|X) = \frac{1}{f(X)} \sum_s \left\{ \left[\sum_k \omega_k f(x_h|\omega^{(h)} = \omega_k) f(\omega_k|\eta_s) \right] \times \prod_{j \neq h} \left[\sum_{k'} f(x_j|\omega^{(j)} = \omega_{k'}) f(\omega_{k'}|\eta_s) \right] f(\eta_s) \right\}. \tag{6}$$

M2a involves four parameters: p_0, p_1, ω_0, and ω_s. We use the priors $\omega_0 \sim U(0, 1)$ and $\omega_s \sim U(1, 11)$, and for each parameter, we use the midpoint of each interval to represent the density in that interval. Thus, the $U(0, 1)$ density for ω_0 is approximated by 10 values 0.05, 0.15, …, 0.95, each with probability 0.1, while the $U(1, 11)$ density for ω_s is approximated using 10 values 1.5, 2.5, …, 10.5, each with probability 0.1. Parameters p_0, p_1, and $p_2 = 1 - p_0 - p_1$ are assumed to have a Dirichlet prior $D(\theta_0, \theta_1, \theta_2)$, as in Huelsenbeck and Dyer (2004), with density

$$f(p_0, p_1, p_2|\theta_0, \theta_1, \theta_2) = \frac{\Gamma(\theta_0 + \theta_1 + \theta_2)}{\Gamma(\theta_0)\Gamma(\theta_1)\Gamma(\theta_2)} p_0^{\theta_0 - 1} p_1^{\theta_1 - 1} p_2^{\theta_2 - 1}. \tag{7}$$

The p_0-p_1-p_2 space is represented by a triangle shown in figure 1. We partition it into $d^2 = 100$ equal-sized triangles and use the center of each to represent the density mass on that triangle (fig. 1). Let the d^2 triangles be labeled 0, 1, …, $d^2 - 1$, starting from the one on the top row, then three on the second row, five on the third row, and finally $2d - 1$ on the last row, row $d - 1$. The mth triangle is on the ith row and jth column ($i = 0, 1, …, d - 1$; $j = 0, 1, …, 2i$), with

$$i = [\sqrt{m}], j = m - i^2, \tag{8}$$

where $[a]$ is the integer part of a. The center of this triangle is at

$$p_0 = \frac{1 + [j/2] \times 3 + (j \bmod 2)}{3d}, \quad p_1 = \frac{1 + (d - 1 - i) \times 3 + (j \bmod 2)}{3d}, \tag{9}$$

where (j mod 2) is the remainder when j is divided by 2.

We use $\theta_0 = \theta_1 = \theta_2 = 1$ in the prior, so that each of the 100 points in the ternary graph of figure 1 receives a prior probability of 0.01. In sum, the 4-D integrals over η in equations (2) and (3) are approximated by a sum over 10^4 points on the 4-D grid in equations (4) and (5).

Under M8, we use $d = 10$ categories for each of the four parameters p_0, p, q, ω_s, and use the midpoint of each interval to represent the density in that interval. We assume the following priors: $p_0 \sim U(0, 1)$, $p \sim U(0, 2)$, $q \sim U(0, 2)$, and $\omega_s \sim U(1,11)$. Thus, p_0 takes any of the 10 values 0.05, 0.15, …, 0.95 with a prior probability of 0.1, each of p and q takes any of the 10 values 0.1, 0.3, …, 1.9 with a prior probability 0.1, while ω_s takes any of the 10 values 1.5, 2.5, …, 10.5 with a prior probability 0.1. To save computation, the beta distribution (for given values of p and q) is discretized using $d = 10$ equally spaced categories, unlike Yang et al. (2000), who used 10 equal-probability categories; that is, beta(p, q) is approximated using 10 categories represented by $\omega = 0.05, 0.15, …, 0.95$, with the proportion for each category equal to the probability mass within that category. Thus, different beta distributions specified by different values of p and q on the grid are represented by the same set of ω values, and $f(x_h|\omega)$ is calculated for the same set of ω values for all sites. This strategy makes the computation feasible (see below), although it may not be as good as the equal-probability scheme for approximating a skewed beta density.

The posterior distribution of parameters η (that is, p_0, p_1, ω_0, ω_2 under M2a and p_0, p, q, ω_s under M8) is

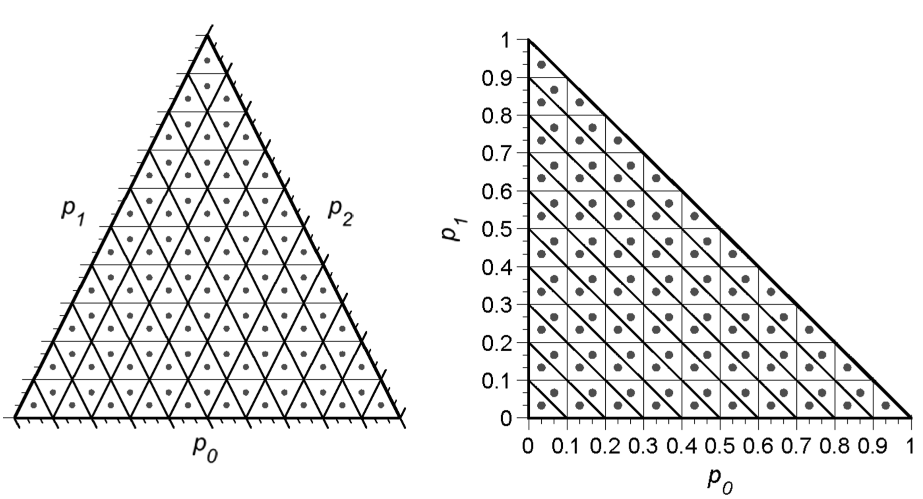

FIG. 1.—Discretization of the Dirichlet prior density for parameters p_0, p_1, and p_2 (= $1 - p_0 - p_1$) under M2a. The parameter space formed by p_0, p_1, and p_2 is a triangle, and this is partitioned into $d^2 = 100$ equal-sized triangles. Each small triangle is represented by a point mass on its center, with the probability on the point mass to be the total density mass on that triangle. Note that $2d - 1 = 19$ distinct values are taken by each of p_0 and p_1 over the 100 points.

calculated as posterior probabilities for the 10^4 points on the 4-D grid. We summarize the distribution by the marginal densities of the parameters. For example, the posterior probability for the proportion parameter p_0 under M8 is

$$\Pr(p_0 = p_0^{(j)}|X) = \frac{1}{f(X)}\sum_s \Big\{ I(p_0 = p_0^{(j)}|\eta_s) \times \prod_{h=1}^{n}\Big[\sum_k f(x_h|\omega^{(h)} = \omega_k) f(\omega_k|\eta_s)\Big] f(\eta_s)\Big\}, \qquad (10)$$

where $p_0^{(j)} = 0.05, 0.15, \ldots, 0.95$, for $j = 1, 2, \ldots, 10$, are the possible values for p_0 on the 4-D grid, and the indicator function $I\,(p_0 = p_0^{(j)}|\eta_s)$ equals 1 if point s on the 4-D grid specifies $p_0^{(j)}$ as the value for p_0 and 0 otherwise. Marginal posterior probabilities for p, q, ω_s are calculated similarly. Under M2a, we calculate the joint posterior probabilities for p_0 and p_1, i.e., for points on the ternary graph of figure 1, and the marginal posterior probabilities for ω_0 and ω_2.

Computational Issues. The computation required by the old NEB method is equivalent to one calculation of the likelihood function, which is about K times as expensive as under the one-ratio model M0, where K is the number of site classes (Nielsen and Yang 1998). The conditional probability of data at each site h given the site class or the ω ratio, $f(x_h|\omega^{(h)} = \omega_k)$, has to be calculated separately for the K site classes. Similarly, in our implementation of the BEB procedure, we would like to calculate the conditional probability for as few ω values as possible. Thus, we fix the branch lengths at the synonymous sites (i.e., the expected number of synonymous substitutions per codon) at their MLEs. Then we calculate $f(x_h|\omega^{(h)} = \omega_k)$ for $2K + 1 = 21$ different ω values under M2a: 10 values for ω_0, 1 for $\omega_1 = 1$, and 10 values for ω_2; and $2K = 20$ different ω values under M8: 10 values for the ω from the beta and 10 values for ω_s. While the computation is several times more expensive than for the NEB procedure, it is much faster than the ML iteration, which requires many calculations of the likelihood function.

BEB Calculations Under Branch-Site and Clade Models

Yang and Nielsen (2002) implemented two branch-site models, A and B, which allow the ω ratio to vary both among sites and among branches. Positive selection is potentially operating on only some branches, called the foreground branches, while the other (background) branches are under purifying selection. The models assume four site classes. In site class 0, all lineages are under purifying selection with a small d_N/d_S ratio ω_0. In site class 1, all lineages are undergoing weak purifying selection or neutral evolution with ω_1 close to 1. In site classes 2a and 2b, a proportion of class-0 and class-1 sites become under positive selection with $\omega_2 > 1$ on the foreground lineages. Model A fixes $\omega_0 = 0$ and $\omega_1 = 1$, while model B estimates those two parameters from the data. Real data analysis suggests that model A is very unrealistic as it fails to account for conserved sites with $0 < \omega < 1$. Thus, we modify model A so that $0 < \omega_0 < 1$ is estimated. We still fix $\omega_1 = 1$ to avoid misclassifying sites under weak purifying selection (with ω close to but less than 1) as positive selection sites. The modified model is still referred to as model A (table 1) and involves four parameters: $\eta = (p_0, p_1, \omega_0, \omega_2)$. Model A is the alternative model and can be used to construct two LRTs. The null model in test 1 is M1a, which assumes two site classes in proportions p_0 and $p_1 = 1 - p_0$ with ratios ω_0 and $\omega_1 = 1$. The null model in test 2 is the same as model A (table 1) except that $\omega_2 = 1$ is fixed. Test 1 may mistake relaxed purifying selection on the foreground branches as positive selection, while test 2 appears to be a direct test of positive selection. A simulation study evaluating the two tests will be reported elsewhere.

Table 1
Parameters in Branch-Site Model A

Site Class	Proportion	Background ω	Foreground ω	Number of Classes in BEB Calculation[a]
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	10
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	1
2a	$(1 - p_0 - p_1)$ $p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$	10 × 10
2b	$(1 - p_0 - p_1)$ $p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$	1 × 10

[a] Number of times that $f(x_h|\omega^{(h)})$ has to be calculated in the BEB algorithm.

Here, we describe our implementation of the BEB procedure for calculating posterior probabilities for site classes under branch-site model A (table 1). As under the site model, we fix the branch lengths at the synonymous sites at their MLEs and accommodate sampling errors in parameters in the ω distribution: $\eta = \{p_0, p_1, \omega_0, \omega_2\}$. We assign a prior $f(\eta)$ and integrate over it. We assume uniform priors $\omega_0 \sim U(0, 1)$ and $\omega_2 \sim U(1, 11)$, in each case using 10 categories to approximate the continuous densities. The prior for parameters p_0 and p_1 is the Dirichlet $D(1, 1, 1)$, and we assign a prior probability of 0.01 for each of the 100 points in the ternary graph of figure 1. The theory is very similar to that under the site models. Similarly, calculation of the probability of the data at each site given the site class and the foreground and background ω ratios, that is, the term equivalent to $f(x_h|\omega^{(h)})$ in equations (1–6), is expensive on large trees. In our implementation, this is calculated for 10 sets of ω ratios for site class 0, 1 set for site class 1, 100 sets for site class 2a, and 10 sets for site class 2b, with 121 sets in total. The rest of the computation does not depend on the size of the tree. We sum over the posterior probabilities for site classes 2a and 2b to obtain the posterior probability that the site is under positive selection along the foreground branches. We also calculated the marginal posterior distributions of the four parameters p_0, p_1, ω_0, and ω_2.

Bielawski and Yang (2004) (see also Forsberg and Christiansen 2003) implemented two clade models, called C and D, to detect divergent selective pressures between clades. Branches in the phylogeny are assumed to fall into two clades. Three site classes are assumed in the models. In site class 0, all lineages are under purifying selection with a small ratio ω_0. In site class 1, all lineages are evolving neutrally or under weak purifying selection with ω_1 close to 1. In site class 2, branches in the two clades are evolving with

Table 2
Parameters in Clade Model C

Site Class	Proportion	ω for Clade 1	ω for Clade 2	Number of Classes in BEB Calculation[a]
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	10
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	1
2	$p_2 = 1 - p_0 - p_1$	ω_2	ω_3	10 × 10

[a] Number of times that $f(x_h|\omega^{(h)})$ has to be calculated in the BEB algorithm.

ω_2 and ω_3, respectively. No positive selection is assumed for either clade; instead the clades may be evolving under divergent selective pressures at some sites. In model C, $\omega_0 = 0$ and $\omega_1 = 1$ are fixed while ω_2 and ω_3 are estimated together with the proportions p_0 and p_1. In model D all parameters are estimated from the data including ω_0 and ω_1. We modify model C so that $0 < \omega_0 < 1$ is estimated while $\omega_1 = 1$ is fixed (table 2) and will refer to the modified model still as model C. The model has five parameters: $\eta = (p_0, p_1, \omega_0, \omega_2, \omega_3)$. The model can be compared with the site model M1a to construct a LRT. Here, we briefly describe our implementation of the BEB method for calculating posterior probabilities for site classes under model C.

We fix all branch lengths measured by the silent rate at their MLEs and accommodate uncertainties in parameters η by assigning a prior and integrating over it. For the prior for p_0 and p_1, we assign equal-probability 0.01 for each of the 100 points in the ternary graph of figure 1. We use $U(0, 1)$ as the prior for ω_0 and $U(0, 3)$ as the prior for ω_2 and ω_3, in each case using 10 categories to approximate the continuous densities. As under the site models, we approximate the 5-D integral using a 5-D grid. Thus, we have to calculate the probability of data at any site given the site class and ω ratios for 111 sets of ω ratios (table 2). Besides posterior probabilities for site classes at each site, we also calculate the marginal posterior distributions of the five parameters.

Results

Analysis of Real Data Sets

Three data sets are analyzed to compare the old NEB and the new BEB approaches for inferring sites under positive selection. We focus our attention on posterior distributions of the parameters (η) in comparison with sampling errors in the MLEs, on posterior probabilities for sites under positive selection inferred by NEB and BEB, and on the effects of the prior for η and of the number of categories in the 4-D grid on calculation of the posterior probabilities.

Human Class I MHC Alleles

A data set consisting of 192 alleles of the human class I MHC alleles from the A, B, and C loci are analyzed. This data set was compiled and analyzed using the old NEB approach by Yang and Swanson (2002). The sequence length is 270 codons. The tree topology estimated by Yang and Swanson (2002) is used here. The F3x4 model of codon frequencies is used. To save computation, we estimate the branch lengths on the tree under model M0 (one-ratio) and use them as fixed when fitting site models M2a and M8.

Table 3 lists the log-likelihood values and the MLEs of parameters under models M2a and M8. Both models have much higher log-likelihood values than their corresponding null models M1a and M7, providing strong evidence for presence of sites under positive selection (results not shown; see table 2 of Yang and Swanson [2002]). Sites inferred to be under positive selection by the NEB and BEB approaches under the two models are listed as well, with the cutoff posterior probability set at $P_b = 95\%$.

Under M2a, the MLEs and their standard errors (SEs) are $\hat{p}_0 = 0.776 \pm 0.022, \hat{p}_1 = 0.140 \pm 0.025, \hat{\omega}_0 = 0.058 \pm 0.009$, and $\hat{\omega}_2 = 5.389 \pm 0.361$. The SEs are approximated using the local curvature of the log likelihood. While those SEs do not take the correlations between parameters into

Table 3
Log-likelihood Values and Parameter Estimates for the Class I MHC Alleles

Model Code	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0 (one-ratio)	1	−8,225.15	$\hat{\omega}=0.612$	None
M2a (PositiveSelection)	4	−7,231.15	$\hat{p}_0=0.776, \hat{p}_1=0.140 (\hat{p}_2=0.084)$, $\hat{\omega}_0=0.058\ (\hat{\omega}_1=1), \hat{\omega}_2=5.389$	**9F**, **24A**, **45M**, **62G**, **63E**, **67V**, 70H, **71S**, **77D**, **80T**, **81L**, 82R, 94T, **95V**, **97R**, 99Y, 113Y, **114H**, **116Y**, **151H**, **152V**, **156L**, **163T**, **167W**
M8 (beta&ω)	5	−7,238.01	$\hat{p}_0=0.915 (\hat{p}_1=0.085)$, $\hat{p}=0.167, \hat{q}=0.717, \hat{\omega}_s=5.079$	**9F**, **24A**, **45M**, **63E**, **67V**, 69A, **70H**, **71S**, **77D**, **80T**, **81L**, **82R**, **94T**, **95V**, **97R**, 99Y, **113Y**, **114H**, **116Y**, **151H**, **152V**, **156L**, **163T**, **167W**

NOTE.—p is the number of parameters in the ω distribution. Branch lengths are fixed at their MLEs under M0 (one-ratio). Estimates of κ range from 1.5 to 1.8. Positive selection sites are inferred at $P_b = 95\%$ with those reaching 99% shown in bold. The lists of sites are identical between NEB and BEB. The reference sequence is from the PDB structure file 1AKJ.

account, their small values suggest that the parameters are reliably estimated in this large data set. In the BEB analysis, the posterior density of p_0 and p_1 is concentrated on two points in the ternary graph of figure 1: $p_0 = 0.73$, $p_1 = 0.13$ and $p_0 = 0.77$, $p_1 = 0.17$, each point receiving a probability of about 0.5. Those points are the central values of two contiguous triangles in the ternary graph of figure 1, indicating that the mode of the joint posterior density for p_0 and p_1 is near the two points. The distributions of ω_0 and ω_2 are concentrated on 0.05 and 5.5, respectively, each with probability ~1.0. Those values may be considered approximate maximum a posteriori estimates and agree well with the MLEs, and the high posterior probabilities reflect the high information content in the data.

The posterior probabilities for the three site classes under M2a are almost identical between NEB and BEB. Those under BEB are often (but not always) less extreme (away from 0 or 1) than those under NEB. Note that the posterior probabilities sum to 1 over the site classes in each method, and less extreme probabilities mean less confidence in inference. The lists of sites under positive selection at the cutoffs $P_b = 95\%$ and 99% are exactly the same between the two approaches (table 3).

Under M8, the results of table 3 are slightly different from those of Yang and Swanson (2002) due to minor differences in the branch lengths used. The MLEs of parameters and their approximate SEs are $\hat{p}_0 = 0.915 \pm 0.007$, $\hat{p} = 0.167 \pm 0.033$, $\hat{q} = 0.717 + 0.163$, and $\hat{\omega}_s = 5.079 + 0.374$. In the BEB analysis, the posterior distribution of p_0 is concentrated on 0.85 and 0.95, with probabilities 0.876 and 0.124. The beta parameter p is concentrated on 0.1, with probability ~1 while q is concentrated on points 0.5 and 0.7, with probabilities 0.810 and 0.179, respectively. Parameter ω_s is concentrated on 4.5 and 5.5, with probabilities 0.135 and 0.865. Again, these approximate posterior estimates agree well with the MLEs. The estimates of beta parameters p and q are slightly more different between the two methods because of different discretization schemes used.

M8 assumes 11 site classes: 10 classes for the beta distribution and 1 class for positively selected sites. Because the old NEB used 10 equal-probability categories to approximate the beta distribution and the new BEB used 10 equally spaced categories, the posterior probabilities for the first 10 site classes are not directly comparable between the two approaches. The posterior probabilities for the positive selection class are very similar. The lists of positive selection sites at $P_b = 95\%$ and 99% are almost identical; the only differences are at sites 82R and 94T, for which $P = 0.992$ and 0.992 by NEB while $P = 0.985$ and 0.987 by BEB.

We also conducted a robustness analysis under M8. First, we used $d = 20$ categories in the BEB calculation, with 160,000 instead of 10,000 points on the 4-D grid, to examine the effect of the number of categories d on calculation of the posterior probabilities. The beta distribution is discretized using $d = 20$ equally spaced categories as well. The posterior probabilities are very similar to those obtained using $d = 10$ categories, when two consecutive categories for $d = 20$ are merged into one category for $d = 10$. Exactly the same sites are inferred to be under positive selection at the 95% and 99% cutoffs for the two values of d. Also, the correlation coefficients in the posterior mean ω are all greater than 0.999 among the three analyses: the old NEB with 10 categories, and the new BEB with 10 or 20 categories. Ten categories appear to be sufficient for discretizing the integral over parameters η.

Next we examine the effect of the prior for η. We applied a triangle prior for p_0 under M8 with density $f(p_0) = 2p_0$, $0 < p_0 < 1$. This prior places more density mass on p_0 close to 1; the prior probabilities for the 10 values 0.05, 0.15, ..., 0.95 are 0.01, 0.03, ..., 0.19. The lists of sites inferred to be under positive selection at $P_b = 95\%$ and 99% are essentially identical to those obtained under the uniform prior (table 3); the only difference is that site 94T had $P = 0.992$ for the uniform prior and 0.987 for the triangle prior. We also used $U(1, 21)$ instead of $U(1, 11)$ as the prior for ω_s. This change had a slightly greater effect. For example, the posterior probabilities for sites 82R, 94T, and 113Y changed from 0.985, 0.987, and 0.993 under the old prior to 0.933, 0.947, and 0.979 under the new prior. Overall, the priors for parameters η had minimal effects on the calculation of the posterior probabilities in this data set.

HIV env *Gene*

The second data set consists of the HIV-1 *env* gene V3 region from 13 HIV-1 isolates, previously analyzed by Yang et al. (2000). The sequence has 91 codons. The F3x4 model of codon frequencies is used. This was intended to be a small data set, suitable for demonstrating differences between the NEB and BEB approaches, but it failed to do so (see below). To see the effects of sequence sampling, we also analyzed a smaller data set of only the first four sequences (accession numbers U68496–U68499).

The MLEs of parameters under models M2a and M8 are listed in table 4, together with the sites inferred to be under positive selection by the NEB and BEB approaches at $P_b = 95\%$. Both M2a and M8 have much higher likelihood values than their corresponding null models M1a and M7, so the LRTs suggest presence of sites under positive selection. Both models identified three sites under positive selection by the old NEB approach.

Under M2a, the MLEs and SEs are $\hat{p}_0 = 0.377 \pm 0.132$, $\hat{p}_1 = 0.441 \pm 0.161$, $\hat{\omega}_0 = 0.060 \pm 0.108$, and $\hat{\omega}_s = 3.626 \pm 0.951$. The large SEs reflect considerable uncertainties in the MLEs. The posterior distribution of p_0 and p_1 has a wide spread around the peak at $p_0 = 0.37$, $p_1 = 0.47$, which has probability 0.098, in comparison with the prior probability 0.01 (see fig. 1). The posterior distribution of ω_0 peaks at 0.05 (with probability 0.4) while that of ω_2 peaks at 3.5 (with probability 0.5). These approximate Bayesian estimates agree well with the MLEs, but their associated small probabilities indicate large sampling errors in the parameters. The posterior probabilities for the three site classes under M2a are similar between NEB and BEB. At $P_b = 95\%$, both approaches identified 28T, 66E, and 87V as sites under positive selection (table 4). At $P_b = 90\%$, site 26N is selected by both approaches as well.

Under M8, the MLEs of parameters and their approximate SEs are $\hat{p}_0 = 0.800 \pm 0.103$, $\hat{p} = 0.167 \pm 0.302$,

Table 4
Log-likelihood Values and Parameter Estimates for the HIV-1 *env* V3 Regions (13 Sequences)

				Positively Selected Sites	
Model Code	p	ℓ	Estimates of Parameters	NEB	BEB
M0 (one-ratio)	1	−1,137.69	$\hat{\omega}=0.901$	Not allowed	Not allowed
M1a (NearlyNeutral)	2	−1,114.64	$\hat{p}_0=0.484(\hat{p}_1=0.516), \hat{\omega}_0=0.079(\omega_1=1)$	Not allowed	Not allowed
M2a (PositiveSelection)	4	−1,106.45	$\hat{p}_0=0.377, \hat{p}_1=0.441(\hat{p}=0.181),$ $\hat{\omega}_0=0.060(\omega_1=1), \hat{\omega}_2=3.626$	**28T**, **66E**, 87V	**28T**, **66E**, 87V
M7 (beta)	2	−1,115.40	$\hat{p}=0.148, \hat{q}=0.118$	Not allowed	Not allowed
M8 (beta&ω)	5	−1,106.39	$\hat{p}_0=0.800(\hat{p}_1=0.200), \hat{p}=0.167,$ $\hat{q}=0.149, \hat{\omega}=3.470$	**28T**, **66E**, 87V	26N, **28T**, 51I, **66E**, **87V**

NOTE.—p is the number of parameters in the ω distribution. Estimates of κ range from 2.4 to 2.8. Positive selection sites are inferred at $P_b = 95\%$ with those reaching 99% shown in bold. The reference sequence is U68496.

$\hat{q} = 0.149 \pm 0.349$,and $\hat{\omega} = 3.470 \pm 1.009$. The large sampling errors, especially for the beta parameters p and q, indicate considerable uncertainty in the MLEs. The marginal posterior densities of p_0, p, q, and ω_s peak at 0.75, 0.70, 1.9, and 3.5, with probabilities at the peaks to be 0.494, 0.181, 0.121, and 0.510, respectively. The approximate Bayesian estimates of p_0 and ω_s agree well with the MLEs, but the estimates of p and q do not because of the different schemes used to discretize the beta distribution. Both analyses suggest that p and q are more poorly estimated than p_0 and ω_s. The posterior probabilities for the positive-selection class are found to be quite similar between NEB and BEB. For example, the probabilities for site 9S are 0.759 and 0.859, with posterior mean ω to be 2.858 and 2.819, for NEB and BEB, respectively. At $P_b = 95\%$, NEB identified 28T, 66E, and 87V to be under positive selection, while at $P_b = 90\%$, sites 26N and 51I are identified as well. All of these five sites reached the 95% cutoff by the BEB approach, which identified two additional sites at $P_b = 90\%$: 69N and 83V. It is interesting that BEB inferred more sites under positive selection than NEB in this data set (table 4).

We also applied the triangle prior for p_0 under M8 with density $f(p_0) = 2p_0$, $0 < p_0 < 1$. This had very minor effect on the posterior distributions of the parameters or on the posterior probabilities of sites under positive selection. The lists of sites under positive selection at $P_b = 0.95$ and 0.99 are identical between the two priors, with almost identical probabilities.

Overall the NEB and BEB approaches produced similar inferences of sites under positive selection for this data set, despite the considerable uncertainties in the MLEs of model parameters. To explore further the differences between the two approaches and to examine the effects of sequence sampling, we analyzed a smaller data set consisting of only the first four sequences (accession numbers U68496–U68499).

In this small data set, the LRT statistic is $2\Delta\ell = 4.1$ for both the M1a-M2a and M7-M8 comparisons, and the null hypotheses are not rejected. Model M2a produced the following MLEs: $\hat{p}_0 = 0.814, \hat{p}_1 = 0.000, \hat{\omega}_0 = 0.866$, and $\hat{\omega}_2 = 6.858$, with very large sampling errors. At $P_b = 0.95$, both NEB and BEB identified one site (28T) to be under positive selection, with $P = 0.96$ for NEB and 0.95 for BEB. Model M8 gave parameter estimates $\hat{p}_0 = 0.815, \hat{p} = 99$, $\hat{q} = 15.2$, and $\hat{\omega} = 6.863$, again with large sampling errors. At $P_b = 0.95$, both NEB and BEB identified site 28T to be under positive selection, with $P = 0.96$ and 0.98 for NEB and BEB, respectively. Overall, NEB and BEB are similar and models M2a and M8 are consistent in this small data set. Note that the single site (28T) identified in this small data set was also identified in the larger 13-sequence data set. The larger data set provided stronger evidence for positive selection in identifying more sites with higher posterior probabilities. A similar pattern was reported for two MHC data sets, with 6 and 192 sequences, respectively, by Swanson et al. (2001) and Yang and Swanson (2002).

HTLV-I tax *Gene*

Twenty sequences of the *tax* gene from the HTLV-I are retrieved from GenBank and analyzed on a star phylogeny, following Suzuki and Nei (2004). The sequences, 181 codons long, are very similar and all differences are singletons. Ancestral sequence reconstruction suggests a total of 23 single-nucleotide mutations: 2 synonymous transitions (at sites 33L and 38E), 19 nonsynonymous transitions (at sites 4P, 39D, 43I, 53V, 60S, 62L, 81G, 85I, 92D, 101S, 108K, 115H, 146S, 152K, 154A, 157N, 161P, 166G, 181V), and 2 nonsynonymous transversions (2C, 69L). Site numbering here refers to sequence AB045401. We use the F3x4 model to accommodate biased codon usage. Application of model M0 (one-ratio) leads to the estimates $\hat{\kappa} = 23.3$ and $\hat{\omega} = 4.87$, with the log-likelihood $\ell = -892.02$. M0 can be compared with the null model that fixes $\omega = 1$. This LRT rejects the null model, with $P = 0.008$. Thus, the average ω across the whole sequence and across all branches on the tree is significantly greater than 1, and there seems to be no doubt that positive selection drives the evolution of the *tax* gene.

We fix the branch lengths at the estimates obtained under M0 when applying the site models. The MLEs under both M2a and M8 are reduced to those under M0, with all sites having $\hat{\omega} = 4.87$ ($\hat{p}_2 = 1$ and $\hat{\omega}_2 = 4.87$ under M2a, and $\hat{p}_1 = 1$ and $\hat{\omega}_s = 4.87$ under M8). The MLEs under the null models M1a and M7 are reduced to $\omega = 1$, with $\ell = -895.50$. The test statistic for the two LRTs comparing M2a with M1a and comparing M8 with M7 is $2\Delta\ell = 6.96$, and the null models are rejected with a marginal P value ~0.03. NEB calculation of posterior probabilities for site classes using such MLEs led to the conclusion that all sites, including the 158 invariant sites, are under positive selection with $P = 1$, as reported by Suzuki and Nei (2004).

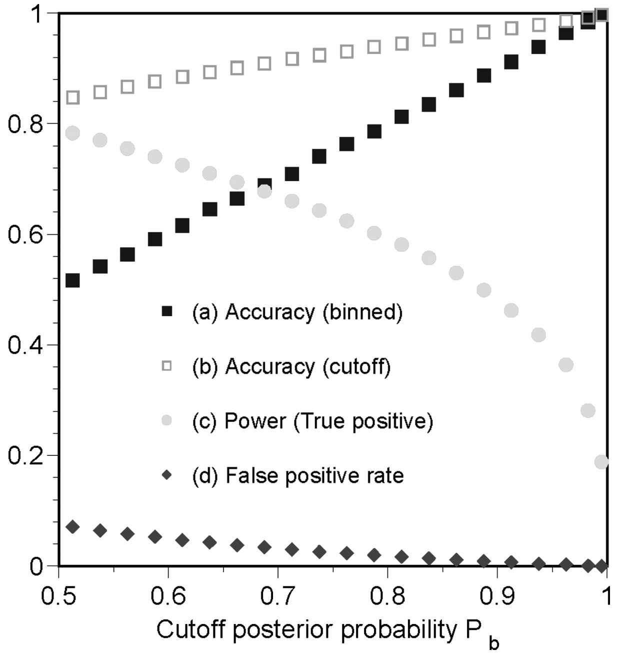

FIG. 2.—A simulation study to illustrate different measures of performance of methods for detecting positive selection sites. See text for simulation conditions. (*a*) Accuracy (binned) is the proportion of sites truly under positive selection among sites with posterior probability P lying in a bin. Sites with $P > 0.5$ are grouped into 21 bins, and Accuracy within each bin is plotted against the midvalue of the bin. Because the correct model and prior are used in analysis, Accuracy equals the posterior probability P. (*b*) Accuracy (cutoff) is the accuracy for all sites exceeding a cutoff probability. (*c*) Power is defined as the proportion of sites inferred correctly to be under positive selection among all true-positive selection sites. (*d*) False-positive rate is the proportion of sites inferred falsely to be under positive selection among all sites not under positive selection.

The BEB approach is then applied to the same data. Under M2a, the posterior density for p_0 and p_1 is the highest at three points (0.033, 0.033), (0.067, 0.067), and (0.033, 0.133), each receiving probability 0.035, compared with the prior probability 0.01. The posterior distribution of ω_0 peaks at 0.95, with probability 0.113, while that of ω_2 peaks at 6.5 and 7.5, each with probability 0.184. Those probabilities are not very different from the prior probability 0.1, indicating lack of information in the data. The 21 sites with a nonsynonymous mutation are inferred to be under positive selection with $0.91 < P < 0.93$, while all other sites are under positive selection with $0.55 < P < 0.61$.

Under M8, the posterior densities for p_0, p, q, and ω_s peak at 0.05, 1.9, 0.1, and 5.5, with probabilities 0.206, 0.106, 0.111, and 0.211, compared with the prior probability 0.1 each. The 21 sites with a nonsynonymous mutation are inferred to be under positive selection with $0.96 < P < 0.97$, while all other sites are inferred to be under positive selection with $0.69 < P < 0.73$. We also applied the triangle prior for p_0 under M8. Under this prior, the posterior densities for p_0, p, q, and ω_s peak at 0.05, 1.9, 0.1, and 5.5, respectively, as under the uniform prior, and the probabilities at the peaks are 0.197, 0.106, 0.111, and 0.207, similar to those under the uniform prior. The posterior probabilities for site classes are identical between the two priors at the level of accuracy used here.

Considering the LRTs and the BEB calculations of posterior probabilities, we suggest that positive selection has affected the evolution of the *tax* gene. The nonsynonymous substitutions seen in the data are likely due to positive selection, although the evidence is marginal.

Computer Simulation

Before describing our simulation experiment, we illustrate the concept of posterior probabilities as well as three performance measures of methods for detecting positive selection sites. Figure 2 shows the results obtained by simulating the data under the prior and analyzing them under the correct prior and model. The tree used is (((A:0.1, B:0.2):0.12, C:0.3):0.123, D:0.4, E:0.5), where the branch length is measured by the expected number of nucleotide substitutions per codon. It is assumed that there is no transition-transversion rate difference so that $\kappa = 1$ and each codon has the equilibrium frequency 1/61. The sequence length is 1,000 codons. The data are generated under M2a with the priors $(p_0, p_1, p_2) \sim D(1, 1, 1)$, $\omega_0 \sim U(0, 1)$ and $\omega_s \sim U(1, 11)$. Each replicate data set is generated by drawing parameters p_0, p_1, ω_0, and ω_s from those priors and then evolving sequences on the tree.

The correct model M2a is used in the analysis, and the true branch lengths are used as fixed. In this case, the posterior probability that a site is under positive selection should be the probability that the site is truly under positive selection. We group the posterior probabilities into bins and in each bin calculate the proportion of sites truly under positive selection. The proportion should then match the probability for the bin. For example, among sites for the bin $0.9 < P < 0.925$, 91.2% of them are found to be truly under positive selection (fig. 2*a*). This is a Bayesian measure, called "accuracy" by Anisimova et al. (2002). If we consider sites with $P > P_b$, the probability that the inferred site is correct will be greater than P_b (fig. 2*b*). For example, among sites which achieved posterior probability 0.90 or higher, 97% of them are truly under positive selection (fig. 2*b*). The second measure is the proportion of sites inferred correctly to be under positive selection among all sites truly under positive selection (fig. 2*c*). This was called power by Anisimova et al. (2002) or proportion of true positives by Wong et al. (2004). It is also known as "sensitivity."

A third measure is the false-positive rate (fig. 2*d*), the proportion of sites not under positive selection that are inferred falsely to be under positive selection. This is a frequentist measure, formulating the problem of identifying positive selection sites as one of testing problem, in which the null hypothesis assumes neutral evolution ($\omega = 1$) while the alternative hypothesis assumes positive selection ($\omega > 1$) (Suzuki and Gojobori 1999). In this formulation, the false-positive rate is also the type I error. This measure was used by Suzuki and Gojobori (1999) and Wong et al. (2004). One minus the false-positive rate is also known as "specificity." Note that the Bayesian posterior probability calculation gives the correct accuracy, but not the frequentist false-positive rate. However, many Bayesian methods are known to have good frequentist properties (see, e.g., pp. 92–108; Carlin and Louis 2000). In figure 1*d*, the false-positive rate does not

Table 5
Performance of BEB and NEB (in parentheses) Inferences of Positive Selection Sites

		5 Taxa (Tree A)		30 Taxa (Tree B)	
Simulation Scheme	Test	Proportion of True Positives	Proportion of False Positives	Proportion of True Positives	Proportion of False Positives
Scheme 1: 100 replicates, 100% ω = 1					
Before LRT	M2a	NA	0.00 (0.33)	NA	0.00 (0.28)
	M8	NA	0.00 (0.24)	NA	0.00 (0.29)
After LRT[a]	M2a-M1a	NA	0.00 (0.02)	NA	0.00 (0.00)
	M8-M7	NA	0.00 (0.03)	NA	0.00 (0.00)
Scheme 2a: 100 replicates, 50% ω = 0.5, 50% ω = 1					
Before LRT	M2a	NA	0.00 (0.14)	NA	0.00 (0.13)
	M8	NA	0.01 (0.08)	NA	0.00 (0.02)
After LRT[a]	M2a-M1a	NA	0.00 (0.00)	NA	0.00 (0.00)
	M8-M7	NA	0.00 (0.00)	NA	0.00 (0.00)
Scheme 4: 100 replicates, 50% ω = 1, 50% ω = 1.5					
Before LRT	M2a	0.02(0.45)	0.01 (0.42)	0.03 (0.32)	0.01 (0.28)
	M8	0.09 (0.38)	0.06 (0.36)	0.09 (0.19)	0.05 (0.16)
After LRT[a]	M2a-M1a	0.02 (0.34)	0.01 (0.32)	0.03 (0.29)	0.01 (0.25)
	M8-M7	0.09 (0.28)	0.06 (0.26)	0.09 (0.16)	0.05 (0.14)
Scheme 6: 50 replicates, 45% ω = 0, 45% ω = 1, 10% ω = 5					
Before LRT	M2a	0.19 (0.18)	0.00 (0.00)	0.76 (0.75)	0.00 (0.00)
	M8	0.42 (0.20)	0.01 (0.00)	0.79 (0.76)	0.00 (0.00)
After LRT[a]	M2a-M1a	0.19 (0.18)	0.00 (0.00)	0.76 (0.75)	0.00 (0.00)
	M8-M7	0.42 (0.20)	0.01 (0.00)	0.79 (0.76)	0.00 (0.00)
Scheme 7: 100 replicates, 12 site classes					
Before LRT	M2a	0.16 (0.16)	0.00 (0.00)	0.43 (0.43)	0.00 (0.00)
	M8	0.25 (0.24)	0.00 (0.00)	0.48 (0.47)	0.00 (0.00)
After LRT[a]	M2a-M1a	0.16 (0.16)	0.00 (0.00)	0.43 (0.43)	0.00 (0.00)
	M8-M7	0.25 (0.24)	0.00 (0.00)	0.48 (0.47)	0.00 (0.00)

NOTE.—Positive selection sites are inferred using the cutoff posterior probability $P_b = 0.95$. The proportion of true positives is defined as the number of sites which are correctly classified as positively selected divided by the total number of positive selection sites simulated. The proportion of false positives is defined as the number of sites which are falsely classified as positively selected divided by the total number of sites that are not positively selected (with $\omega \leqslant 1$). The NEB results for schemes 4 and 6 are from Wong et al. (2004). Scheme 7 assumes 12 site classes with ω ratios 0, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8, 1, 2, 3, 4, and 5 in proportions 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05, 0.05, and 0.05, respectively.

[a] The NEB and BEB calculations are applied only if the LRT is significant at the 5% level.

match and is much lower than $1 - P_b$. For example, at the cutoff $P_b = 0.7$, the false-positive rate is only 0.03, much lower than $1 - 0.7 = 0.3$ (fig. 2*d*). Note that Suzuki and Nei (2002) confused the Bayesian posterior probabilities with the frequentist type I error rate when they claimed that $1 - P_b$ should equal the nominal *P* value.

In the following simulation, we examine the frequentist false-positive rate and the power (proportion of true positives) of the new BEB approach in comparison with the old NEB approach for detecting positive selection sites. Data are simulated using fixed values of parameters for the ω distribution. We address two major questions: (1) does BEB overcome the problem of high false positives of NEB in small data sets? (2) does the BEB correction cause a loss of power in large data sets in which NEB was working well? We used two simulation schemes (4 and 6) of Wong et al. (2004) plus a new scheme. Scheme 4 assumes two site classes in proportions 1:1 with ω = 1 and 1.5. The old NEB produced many false positives under this scheme (Wong et al. 2004), and it is interesting to know whether the BEB is an improvement. Scheme 6 assumes three site classes in proportions 45%, 45%, and 10% with ω ratios 0, 1, and 5, respectively. Under this scheme, NEB performed very well (Wong et al. 2004), and it is interesting to know whether the BEB correction causes any loss of power. The third scheme (scheme 7) is new and assumes 12 site classes in proportions 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05 with ω ratios 0, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8, 1, 2, 3, 4, and 5, respectively. This scheme is used to evaluate the robustness of the analysis and to address the concern that both M2a and M8 assume one ω value for all sites under positive selection while those sites may be expected to be under different strengths of selection.

Simulation conditions follow Wong et al. (2004). Two trees with either 5 or 30 taxa are used with fixed branch lengths. The tree length, that is, the expected number of nucleotide substitutions per codon along all branches in the tree, is 3. Again the model assumes no transition-transversion bias ($\kappa = 1$) or codon usage bias ($\pi_j = 1/61$). The sequence length is 500 codons. Data sets were simulated using the evolver program in the PAML package (Yang 1997) and analyzed using the codeml program, which implements both the NEB and BEB approaches. The correct tree topology is used, but the branch lengths are estimated by ML.

The results are summarized in table 5. Under scheme 4, the old NEB has high false-positive rates caused by inaccurate MLEs of parameters in the ω distribution. The BEB procedure corrects for the problem and reduces the false-positive rate considerably. For example, under M2a the false-positive rate is 42% for NEB but only 1% for BEB at the cutoff $P_b = 95\%$. The false-positive rate for BEB under M8 is higher than under M2, at 5% in the large tree and 6% in the small tree. Under this scheme, BEB has very low true-positive rate, never identifying more than 9% of all positive selection sites, even in the large tree. It appears to be very difficult to identify positively selected sites with ω as low as 1.5. The apparent high power of NEB in this scheme is clearly unreliable.

Under schemes 6 and 7, the old NEB performed well, with the false-positive rate at ~0% at the cutoff $P_b = 95\%$. The method also has good power in identifying positive selection sites, especially under scheme 6. The new BEB also performed well, with the false-positive rate at 0%–1% at the cutoff $P_b = 95\%$. BEB never had lower true-positive rate than NEB, and indeed in some cases, it even recovers more positive selection sites than NEB. For example, under scheme 6 and M8 for the small tree, the true-positive rate increased from 19% for NEB to 42% for BEB. The true-positive rate is higher in the 30-taxa tree than in the 5-taxa tree even though the total tree length is the same, probably because the large tree allows the same number of changes to be distributed on many branches, so that the data are more informative (for example, about the codons at ancestral nodes) than on the small tree.

To examine the false-positive rate of the BEB procedure when the data contain no positive selection sites, we also simulated data sets under schemes 1 and 2a of Wong et al. (2004). Scheme 1 assumes that all sites have ω = 1 (corresponding to a pseudogene), while scheme 2a assumes that 50% of sites have ω = 0 and 50% have ω = 1. Schemes 1 and 2a are similar to schemes 4 and 6 except for the absence of sites under positive selection. It is not possible to calculate true-positive rates as there are no true-positive sites. The false-positive rate for BEB at the cutoff $P_b = 0.95$ is found to be 0 for both schemes 1 and 2a, for both trees, and both before and after the LRT. The error rate is 0 even if a less stringent criterion $P_b = 0.5$ is applied. Thus, the false-positive rate for BEB is lower in schemes 1 and 2a, where no positive selection sites are present, than in corresponding schemes 4 and 6, where some sites are under positive selection.

We also note that BEB maintains a low false-positive rate even when the LRT has not been performed first. However, we suggest that to answer the question whether there are any sites in the sequence under positive selection, the LRT should be used, while the BEB should be used to identify positive selection sites when the LRT indicates that such sites exist. Overall, the BEB correction appears to avoid the high false-positive rates of the NEB approach in small noninformative data sets, while it has not caused any loss of power in large informative data sets. It also appears that the BEB procedure tends to be conservative if considered a frequentist test; the false-positive rate is often much lower than $1 - P_b$ when sites are identified at the cutoff posterior probability P_b.

Discussion

In all three real data sets analyzed in this paper, the prior for p_0 under M8 was found to have minimal effects on the posterior distributions of model parameters or on posterior probabilities for site classes. This insensitivity, especially in the small data sets of HIV *env* genes and HTLV-I *rax* genes, appears to be due to the fact that priors on η are second-level priors as far as inference on $\omega^{(h)}$ is concerned. While no robustness analysis has been conducted on all parameters η under M8 or under M2a, one may expect that the pattern is general. We also note that several previous studies demonstrated that test of positive selection and identification of sites under positive selection were insensitive to minor errors in the tree topology or to different estimates of the branch lengths. For example, the tree topology was found to have minimal effects by Suzuki and Gojobori (1999), Yang et al. (2000), and Swanson et al. (2001). Yang (2000) tested a few different ways of estimating branch lengths in the tree, including one using nucleotide-substitution models, and found that they all produced highly similar inferences of positive selection sites. Thus, we expect that our fixation of branch lengths to their MLEs in the BEB calculation should not introduce large errors.

We used three real data sets to evaluate the differences between the NEB and BEB approaches. The two methods are different when the MLEs are extreme as in the HTLV-I *rax* gene. What is striking is perhaps the similarity between the two methods in very small data sets, such as the HIV *env* genes. The real data analysis also suggests that models 2A and 8 usually gave similar conclusions, as found in early studies (e.g., Yang et al. 2000; Swanson et al. 2001). This pattern appears to suggest that previous studies using the NEB approach should be fine as long as the data set is not too small and the estimates are not extreme (say, with estimates of proportions to be 0 or 1). However, if the data consist of few short sequences, or if estimates of ω_s are only slightly larger than 1, it may be worthwhile to use the new BEB method to confirm results. Sequence sampling seems to have greater effects than either the prior for parameters or the different methods (NEB vs. BEB).

The simulation study suggests that the BEB method in general appears to have good statistical properties. In small data sets, the BEB does not have the high false-positive rate of the NEB approach, while in large data sets, the BEB seems at least as powerful as NEB. The BEB appears often to be conservative under the frequentist criterion, with the false-positive rate to be lower than 5% if a cutoff posterior probability of $P_b = 95\%$ is applied.

The extensive simulation studies performed by Anisimova, Bielawski and Yang (2001) and Wong et al. (2004) demonstrate that the LRTs for detecting positive selection, suggested by Nielsen and Yang (1998) and Yang et al. (2000), have good statistical properties over a wide range of conditions. Analyses of both real and simulated data sets in this study suggest that the new BEB method is reliable in both small and large data sets and also has good power for identifying individual positively selected sites, especially in large data sets or with strong selective pressure. Together, those methods provide a robust and trustworthy framework for inference of positive selection affecting protein-coding genes. However, it is important to be aware of the inherent limitations of these methods. First, they have appreciable power to detect positive selection only if multiple substitutions have occurred at the same codon site throughout the phylogeny. If positive selection does not involve recurrent fixations of nonsynonymous mutations at the same sites, those methods may fail. For example, Guindon et al. (2004) demonstrated that in some HIV-1 genes, the selective pressure varies not only among sites but also among lineages. Second, a number of assumptions are made in the tests for positive selection, which may be violated in real data. For example, simulations demonstrated that the LRTs are not robust to frequent intragenic recombinations (Anisimova, Nielsen, and Yang 2003). Likewise, the methods accommodate variable nonsynonymous rates among sites but assume the same synonymous rate, and it is possible that varying mutation rates (or other mutational parameters) among sites may mimic the effect of positive selection (Kosakovsky Pond, Frost, and Muse 2004). We encourage more work identifying cases where the likelihood methods for detecting positive selection might fail. Only by identifying such cases is it possible to further improve the current framework and construct even better statistical methods.

Acknowledgments

We thank Stephane Aris-Brosou, Dave Dale, Stephane Guindon, Konrad Scheffler, and an anonymous referee for constructive comments. This research was supported by grants from the Biotechnological and Biological Sciences Research Council, the Human Frontier Science Program grant (Y0055/2001-M) and by NSF/NIH (DMS/NIGMS-0201037).

Literature Cited

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol. Biol. Evol. **18**:1585–1592.

———. 2002. Accuracy and power of Bayes 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 the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229–1236.

Bielawski, J. P., and Z. Yang. 2001. Positive and negative selection in the DAZ gene family. Mol. Biol. Evol. **18**:523–529.

———. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. **59**:121–132.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA **97**: 5322–5327.

Carlin, B. P. and T. A. Louis. 2000. Bayes and empirical Bayes methods for data analysis. London, Chapman and Hall.

Carlin, B. P., and A. E. Gelfand. 1990. Approaches for empirical Bayes confidence intervals. J. Am. Stat. Assoc. **85**:105–114.

Deely, J. J., and D. V. Lindley. 1981. Bayes empirical Bayes. J. Am. Stat. Assoc. **76**:833–841.

Filip, L. C., and N. I. Mundy. 2004. Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. Mol. Biol. Evol. **21**:1504–1511.

Ford, M. J. 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol. Biol. Evol. **18**:639–647.

Forsberg, R., and F. B. Christiansen. 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol. Biol. Evol. **20**:1252–1259.

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

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

Haydon, D. T., A. D. Bastos, N. J. Knowles, and A. R. Samuel. 2001. Evidence for positive selection in foot-and-mouth-disease virus capsid genes from field isolates. Genetics **157**:7–15.

Huelsenbeck, J. P., and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. J. Mol. Evol. **58**:661–672.

Kosakovsky Pond, S. L., S. D. W. Frost, and S. V. Muse. 2004. HyPhy: hypothesis testing using phylogenies. Bioinformatics (in press).

Laird, N. M., and T. A. Louis. 1987. Empirical Bayes confidence intervals based on bootstrap samples. J. Amer. Stat. Assoc. **82**:739–750.

Lane, R. P., J. Young, T. Newman, and B. J. Trask. 2004. Species specificity in rodent pheromone receptor repertoires. Genome Res. **14**:603–608.

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.

Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12**:219–236.

Mondragon-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res. **12**:1305–1315.

Morris, C. 1983. Parametric empirical Bayes inference: theory and applications. J. Am. Stat. Assoc. **78**:47–65.

Moury, B. 2004. Differential selection of genes of cucumber mosaic virus subgroups. Mol. Biol. Evol. **21**:1602–1611.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

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.

Suzuki, Y., and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16**:1315–1328.

Suzuki, Y., and M. Nei. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **19**:1865–1869.

———. 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. **21**:914–921.

Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro. 2001. Positive Darwinian selection in the evolution of mammalian female reproductive proteins. Proc. Natl. Acad. Sci. USA **98**: 2509–2514.

Takebayashi, N., P. B. Brewer, E. Newbigin, and M. K. Uyenoyama. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. **20**:1778–1794.

Twiddy, S. S., C. H. Woelk, and E. C. Holmes. 2002. Phylogenetic evidence for adaptive evolution of dengue viruses in nature. J. Gen. Virol. **83**:1679–1689.

Wong, W. S. W., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**:1041–1051.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556. (http://abacus.gene.ucl.ac.uk/software/paml.html).

———. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. **51**:423–432.

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

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

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.

Zanotto, P. M., E. G. Kallas, R. F. Souza, and E. C. Holmes. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics **153**:1077–1089.

Arndt von Haeseler, Associate Editor

Accepted January 25, 2005