# Likelihood-Based Clustering (LiBaC) for Codon Models, a Method for Grouping Sites according to Similarities in the Underlying Process of Evolution

*Le Bao,* Hong Gu,* Katherine A. Dunn,† and Joseph P. Bielawski*†*

*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada; and †Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Models of codon evolution are useful for investigating the strength and direction of natural selection via a parameter for the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$). Different codon models are available to account for diversity of the evolutionary patterns among sites. Codon models that specify data partitions as fixed effects allow the most evolutionary diversity among sites but require that site partitions are a priori identifiable. Models that use a parametric distribution to express the variability in the ω ratio across site do not require a priori partitioning of sites, but they permit less among-site diversity in the evolutionary process. Simulation studies presented in this paper indicate that differences among sites in estimates of ω under an overly simplistic analytical model can reflect more than just natural selection pressure. We also find that the classic likelihood ratio tests for positive selection have a high false-positive rate in some situations. In this paper, we developed a new method for assigning codon sites into groups where each group has a different model, and the likelihood over all sites is maximized. The method, called likelihood-based clustering (LiBaC), can be viewed as a generalization of the family of model-based clustering approaches to models of codon evolution. We report the performance of several LiBaC-based methods, and selected alternative methods, over a wide variety of scenarios. We find that LiBaC, under an appropriate model, can provide reliable parameter estimates when the process of evolution is very heterogeneous among groups of sites. Certain types of proteins, such as transmembrane proteins, are expected to exhibit such heterogeneity. A survey of genes encoding transmembrane proteins suggests that overly simplistic models could be leading to false signal for positive selection among such genes. In these cases, LiBaC-based methods offer an important addition to a "toolbox" of methods thereby helping to uncover robust evidence for the action of positive selection.

## Introduction

The ratio of nonsynonymous (amino acid altering) and synonymous (silent) substitution (commonly denoted as the $d_N/d_S$ ratio or ω) has proven to be a valuable index of the strength and direction of natural selection pressure (for a review, see Bielawski and Yang 2004). A value of $\omega < 1$ indicates that, on average, amino acid altering changes to a protein have had negative fitness consequences for the organism, and a value of $\omega > 1$ indicates that amino acid altering changes have increased fitness. However, because the strength and direction of selection on a given amino acid is a function of the 3-dimensional structure of the protein, most proteins are expected to be subject to variable selection pressures among codon sites. Furthermore, the combination of structural and functional constraints means that adaptive changes are expected to occur at only a small subset of sites, with most sites subject to strong negative selection (e.g., Gillespie 1991). For this reason, considerable effort has been devoted to developing methods to infer if individual amino acid sites are subject to positive or negative selection pressure (e.g., Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang and Swanson 2002).

The different methods have unique advantages and limitations, and it is not our purpose to review all previous work; thorough reviews and comparisons are available from several sources (e.g., Wong et al. 2004; Kosakovsky Pond and Frost 2005; Massingham and Goldman 2005; Yang et al. 2005). However, several studies indicate that inadequate modeling of the underlying substitution process can negatively impact estimates of substitution rates (e.g., Yang and Nielsen 2000; Dunn et al. 2001; Aris-Brosou and Bielawski 2006) and classification of sites according to selection pressure (e.g., Anisimova et al. 2002; Wong et al. 2004; Kosakovsky Pond and Muse 2005). Regardless of the method of inference, biological interpretation of any differences among sites in ω requires that such differences are due to selection pressure alone. If several aspects of the substitution process are not constant across sites in real data, estimated differences in ω might not be solely due to differences in selection pressure. Recent work on fixed-effect codon models provides increased flexibility to model heterogeneity among sites in the transition to transversion rate ratio, codon frequencies, relative rates, and selection pressure (Bao et al. 2007). However, these models are not useful when there is no prior information by which to partition the data, and they cannot be fit to partitions comprised of a single site due to the large number of parameters.

The field of statistical clustering offers a possible solution to the problem of resolving groups of sites under similar selection pressure when several aspects of the substitution process vary among sites. We proceed under the assumption that codon sites in a multiple sequence alignment could be comprised of "clusters" that share a common generating model. In this paper, we describe a new clustering method, called likelihood-based clustering (LiBaC), which allows us to maximize the likelihood of the data when different subsets of codon sites have different evolutionary models. Hence, LiBaC can be used to identify sites subject to positive selection when several aspects of the substitution process vary among codon sites. We use computer simulation and real data analyses to evaluate the performance of LiBaC and several alternative approaches. Our simulation studies reveal that estimates of ω can be negatively impacted when other aspects of the substitution process, such as transition to transversion ratio and codon frequencies, are not constant across sites. We

Key words: codon model, likelihood-based clustering, Bayes error rate, nonsynonymous/synonymous rate ratio, positive Darwinian selection.

E-mail: j.bielawski@dal.ca.

*Mol. Biol. Evol.* 25(9):1995–2007. 2008
doi:10.1093/molbev/msn145
Advance Access publication June 26, 2008

introduce the use of the "Bayes error rate" as an objective standard for the performance of methods that classify sites according to selection pressure. Using this framework, we show that performance is expected to depend on the data, with some evolutionary scenarios representing more difficult classification problems than others. We find that LiBaC can provide improved estimates of model parameters under a variety of scenarios and can approach the theoretical upper boundary on classification performance.

## Theory and Methods

### Fixed-Effect Codon Models

We employ the basic codon model of Goldman and Yang (1994), which assumes that the process of substitutions from one codon to another is a Markov process. The *ij*th element $P_{ij}(t)$ in transition matrix $P(t)$ gives the probability going from codon $i$ to codon $j$ during time $t$. There are 64 codons, and the 3 stop codons (UAA, UAG, and UGA) are excluded from the state space of the model because they do not occur within a functional protein-coding gene. The transition matrix $P(t)$ can be calculated by $P(t) = e^{Qt}$, where $Q = \{q_{ij}\}$ is a 61 × 61 rate matrix; the element $q_{ij}$ denotes the instantaneous substitution rate from codon $i$ to codon $j$, and only single-nucleotide substitutions are permitted to occur instantaneously. The elements of $Q$ are parameterized as follows:

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

where $\pi_j$ is the frequency of the $j$th codon, κ is the transition to transversion rate ratio, and ω is the nonsynonymous to synonymous rate ratio.

In the fixed-effect approach, codon sites are partitioned into $G$ different groups, and each group may be permitted to have a different evolutionary model (Yang and Swanson 2002; Kosakovsky Pond and Muse 2005; Bao et al. 2007). Thus, these models can accommodate considerable among-site variability in the substitution process; any combination of heterogeneity or homogeneity among site groups for the parameters κ, ω, and π may be specified. Typically, site groupings are derived from a priori knowledge of a protein's structural and functional domains; for example, buried versus exposed sites in the 3-dimensional structure of the protein. Two problems have hindered the application of fixed-effect models: 1) for many genes, there is no obvious a priori criterion by which sites can be partitioned into different groups and 2) there is often uncertainty about how to partition sites even when some a priori information is available. In the latter case, it is often possible to divide sites in many different ways according to different criteria (e.g., active, internal, surface, highly variable, or conserved sites), all of which might serve as a suboptimal basis to partition sites.

### Likelihood-Based Clustering

We develop a LiBaC method to partition codon sites into groups where each group has a different model, and the likelihood over all sites is maximized (LiBaC). Given a tree topology, the likelihood of the observed data at the $i$th codon site is $f(x_i|\theta)$. Let $f_k(x_i|\theta_k)$ be the probability of observing codons at site $i$ under the hypothesis that this codon site belongs to the $k$th cluster (i.e., group of sites), where $\theta_k$ is the collection of parameters corresponding to the $k$th cluster ($k = 1, \ldots, G$). Suppose the mixing probabilities are $\tau_1, \tau_2, \ldots \tau_G$ such that $P(x_i$ belongs to the $k$th cluster$) = \tau_k$. The purpose is then to maximize the mixture log likelihood

$$\ln P(X|\theta_k, \tau_k, k = 1, \ldots G) = \sum_{i=1}^{n} \ln\left\{ \sum_{k=1}^{G} \tau_k f_k(x_i|\theta_k) \right\}.$$

A difficulty results from the fact that a summation appears inside the logarithm. The typical algorithm to solve this problem is the expectation–maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). The EM algorithm augments the observed data $x_i$ by a $G$-dimensional binary random variable (latent variable) $z_i = (z_{i1}, \ldots, z_{iG})$ with a particular element $z_{ik} = 1$ if $x_i$ belongs to the $k$th cluster; $z_{ik} = 0$ otherwise. The marginal distribution over $z_i$ is specified according to the mixing coefficients $\tau_1, \tau_2, \ldots \tau_G$. The density of an observation $x_i$ given $z_i$ is given by $\prod_{k=1}^{G} f_k(x_i|\theta_k)^{z_{ik}}$; thus, the joint probability of the so-called "complete" data $\{x_i, z_i\}$ can be written as $\prod_{k=1}^{G} [\tau_k f_k(x_i|\theta_k)]^{z_{ik}}$. The EM algorithm circumvents the difficulty of maximizing the mixture log likelihood by maximizing the joint log likelihood of the complete data. If the value of $z_i$ is given, then the joint probability of the complete data $\{x_i, z_i\}$ simply takes the form of $\tau_k f_k(x_i|\theta_k)$. However, as $z_i$ is unknown in practice, the joint probability of the complete data is estimated by its expected value under the posterior distribution of the latent variable. The resulting expected complete data log likelihood is

$$l(\theta, \theta^{\text{old}}) = \sum_{i=1}^{n} \sum_{k=1}^{G} P\left(z_{ik} = 1 \middle| x, \theta^{\text{old}}\right) \log\left[\tau_k f_k\left(x_i \middle| \theta_k\right)\right].$$

Note that independence among the codon sites of a gene is assumed; hence, the log likelihood of the whole sequence is simply the sum of the log likelihood of each site as above (Goldman and Yang 1994). The E-step of the EM algorithm evaluates the posterior distribution of the latent variable for given parameter values $\theta^{\text{old}}$. The M-step estimates the $\theta^{\text{new}}$ by maximizing the above complete data log likelihood. Each pair of successive E- and M-steps gives rise to a revised $\theta^{\text{new}}$ and increases the log-likelihood value. The convergence can be checked by either the log likelihood or the parameter values.

For the Markov models of codon evolution, the M-step proves to be very time consuming; an alternative procedure which improves this point by compromising the values in E-step is also developed. More specifically, the posterior probabilities in E-step $P(z_{ik}=1|x,\theta^{old})$ will be replaced by 1 if site $i$ has the highest posterior probability coming from cluster $k$ and $P(z_{ij}=1|x,\theta^{old})=0$ for $j \neq k$. This corresponds to setting a classification step between the E-step and M-step, which was termed as classification EM by Celeux and Govaert (1992). This can also be viewed as a generalization of the popular K-means clustering algorithm (MacQueen 1967) using likelihood as criterion in the codon evolutionary models. This method is hereafter referred as "hard-LiBaC," and the former method, based on an exact EM algorithm, is referred to as "soft-LiBaC." The speed of hard-LiBaC, soft-LiBaC, and several alternatives is provided in the Supplementary Material online for a variety of real data sets.

We summarize both soft-LiBaC and hard-LiBaC procedures as below:

Initial step: Use M0 (Goldman and Yang 1994) to separately estimate the parameters for sites initially placed in a user-defined site partition, or use M3 (Yang et al. 2000) to estimate the parameters for a user-defined number of site classes $G$.

E-step: Based on the current parameter estimates, $\theta^{old}$, compute the posterior probabilities by Bayes' rule:

$$P(z_{ik}=1|x,\theta)=\frac{\tau_k f_k(x_i|\theta_k)}{\sum_{k=1}^{G}\tau_k f_k(x_i|\theta_k)}.$$

C-step (only for hard-LiBaC, skip this step for soft-LiBaC): Set $\max_{k=1,\ldots,G}P(z_{ik}=1|x,\theta)=1$ and all others $P(z_{ij}=1|x,\theta)=0$.

M-step: Reestimate parameters $\theta^{new}=(t's,\pi,\kappa,\omega)$ by maximizing $l(\theta,\theta^{old})$using the current posterior probabilities:

$$w_k=\sum_{i=1}^{n}P(z_{ik}=1|x,\theta),$$
$$\tau_k=w_k/n,$$
$$l(\theta,\theta^{old})=\sum_{i=1}^{n}\sum_{k=1}^{G}P(z_{ik}=1|x,\theta^{old})\log[\tau_k f_k(x_i|\theta_k)].$$

Convergence: Check for convergence of the log likelihood $l(\theta,\theta^{old})$evaluated at $\theta=\theta^{new}$. If the convergence criterion is not satisfied return to E-step.

Because EM is a hill-climbing algorithm, the starting point for LiBaC (i.e., the initial cluster allocations and parameter values of the model) could influence the outcome; depending on the staring point, LiBaC could converge at a suboptimal peak in likelihood.

We implement LiBaC under 2 levels of among-class heterogeneity in the substitution process, referred to as LiBaC1 and LiBaC2. LiBaC1 permits heterogeneity among classes of sites in codon frequencies ($\pi$'s), natural selection pressure ($\omega$), transition to transversion ratio ($\kappa$), and branch length scale parameter ($c$). LiBaC2 differs from LiBaC1 in that codon frequencies ($\pi$'s) are assumed to be homogenous among classes of sites. Class-specific parameter values are estimated via maximum likelihood (ML) except that $\pi$'s are estimated empirically. LiBaC1 and LiBaC2 make the same assumptions as fixed-effect codon models FE1 and FE2 (Yang and Swanson 2002; Bao et al. 2007). However, FE1 and FE2 treat each branch length in a reference data partition as free model parameters and estimate them by ML. As the computation cost of such an approach is prohibitive under LiBaC, we obtain an initial set of branch lengths via ML under M0, which is assumed to be proportionally correct for each class of sites, and a scale parameter, $c_i$, is then estimated for each class. This way we avoid estimating a full set of branch length parameters. Note that the scale parameters are rescaled such that $c_1 = 1$ (i.e., $c_1 = c_1/c_1$, $c_2 = c_2/c_1$, etc.). We found that this treatment provides very similar results to those obtained under FE1 and FE2 (Yang and Swanson 2002; Bao et al. 2007) in both simulated and real data analyses while providing considerable savings in computational time. LiBaC was implemented by modifying the codeml program of the PAML package (Yang 1997); the modified program is available at http://www.bielawski.info or http://www.mathstat.dal.ca/~hgu/publications.html.

## Simulations and Analyses

Simulation is used to evaluate LiBaC (both soft and hard) according to 1) the classification of codon sites into different classes and 2) the reliability of parameter estimates for different groups. Our simulations include a balanced 2-class scenario, 2 unbalanced scenarios based on continuously variable selection pressure among sites, and a benchmark scenario adapted from Anisimova et al. (2001).

To better understand the simulation results, we first need to describe the criteria that we use to evaluate the classification. Three measures of performance are considered; precision (accuracy), recall (power), and misclassification rate. These terminologies are closely associated with the statistics of classification. Precision is the percentage of correctly predicted positive selection sites among the total predicted to be under positive selection. Note that precision is also known as "accuracy" (Anisimova et al. 2002) or the "positive predictive value." Recall is measured as the percentage of correctly predicted positive selection sites among the number of all sites truly subject to positive selection. This measure is also known as "power" (Anisimova et al. 2002) or "sensitivity." Misclassification is the total percentage of sites misclassified from its own group to any other group. These 3 measures are not independent; given any 2 measures, and the sizes of the site groups, the third can be calculated. For example, for 2 groups, the misclassification rate is equal to $\frac{n_1}{n_1+n_2}(1+recall(1-2\,recall)/precision)$, where $n_1$ and $n_2$ are the number of sites in each group. It is not hard to see that with increasing precision, the recall decreases and vice versa.

The concepts of precision (i.e., accuracy) and recall (i.e., power) are often used in the context of measuring the performance of a method at detecting positive selection (Anisimova et al. 2002). As the problem is now formulated as a classification problem, the misclassification rate has its own importance. For any given model, the theoretical lower

bound for misclassification rate, referred to as the Bayes error rate (Fukunaga 1985; Tumer and Ghosh 2003), can be found as an objective standard for comparison of classification methods. The Bayes error rate is defined as the expected misclassification rate, with the expectation taken on the true model. Hence, it is the theoretical lower bound on the misclassification rate. A high Bayes error rate indicates a hard classification problem, and a method that tends to give an average misclassification rate close to the Bayes error rate is regarded as a good method.

Bayes error rates for the models involved here are not analytically available. We thus estimate this theoretical lower bound by simulation. We simulate 100 data sets under 2 generating codon models with sequence lengths of 500 sites. For each site pattern, we calculate the posterior probabilities that a site was generated under either model according to the true model parameters and assign the site to a group according to the higher posterior probability. The average misclassification rate over the 100 data sets is the estimated Bayes error rate.

Similar to the above, we can also calculate the average precision and recall over 100 data sets by assigning the sites to groups using different posterior probability cutoff values calculated under the true generating models. A precision versus recall curve from these data shows the theoretical upper bound of precision for each possible value of the estimated recall. Given a set of candidate methods, those methods that yield values of precision and recall close to this "idealized precision–recall curve" are regarded as optimized methods. Note that a cutoff posterior probability can be chosen that always yields 100% precision (and 0% recall). As this is not an optimal solution to a classification problem, most investigators seek to achieve some compromise between precision and recall. In this study, the average misclassification rate is calculated for a cutoff posterior probability of 50%.

It is not possible to use standard approaches such as likelihood ratio test (LRT) and Akaike information criterion under LiBaC to select the optimal number of site classes ($G$). Although a simulation-based approach can be employed on a case-by-case basis, this would be too time consuming under LiBaC to be of practical value. Our solution to this problem is motivated by a class of techniques called "indirect inference" where an auxiliary criterion is employed when complex likelihood functions are difficult to work with Gouriéroux et al. (1993) and Genton and Ronchetti (2003). In the spirit of indirect inference, we employ the likelihood score obtained under model M3 as an auxiliary criterion from which we test the goodness of fit of different numbers of site classes by using the LRT. Several simulation scenarios were used to verify the performance of this auxiliary criterion; acceptable performance was achieved over a variety of scenarios under $\alpha = 0.01$ (Supplementary Material online). We note that mixture models allowing greater complexity than M3 could serve as better auxiliary criteria (Kosakovsky Pond SL, personal communication), and such models can be implemented by using the program HyPhy (Kosakovsky Pond et al. 2005). A test of the number of site classes by using M3 as the auxiliary criterion is hereafter referred to as the "surrogate test" for number of site classes.

**Table 1**
**Parameter Values for the 8 Scenarios of Simulation Study 1**

| | Model Parameter Values in Site Classes 1 and 2 | | | |
|---|---|---|---|---|
| Scenario | $\omega_1, \omega_2$ | $c_1, c_2$ | $\kappa_1, \kappa_2$ | $\pi_1, \pi_2$ |
| A1 | 0.3, 1.5 | 1, 5 | 1, 5 | 1/61, empirical |
| A2 | 0.3, 1.5 | 1, 5 | 1, 5 | 1/61, 1/61 |
| B1 | 0.3, 1.5 | 1, 5 | 2, 2 | 1/61, empirical |
| B2 | 0.3, 1.5 | 1, 5 | 2, 2 | 1/61, 1/61 |
| C1 | 0.3, 1.5 | 1, 1 | 1, 5 | 1/61, empirical |
| C2 | 0.3, 1.5 | 1, 1 | 1, 5 | 1/61, 1/61 |
| D1 | 0.3, 1.5 | 1, 1 | 2, 2 | 1/61, empirical |
| D2 | 0.3, 1.5 | 1, 1 | 2, 2 | 1/61, 1/61 |

NOTE.—Site classes 1 and 2 were comprise of 500 codon sites. Empirical codon frequencies were obtained from the abalone sperm lysin data set distributed online as part of the PAML package.

### *Simulation Study 1: Balanced, 2-Class Scenarios*

The purpose of this simulation is to evaluate LiBaC under a wide variety of scenarios of among-site heterogeneity. We adopt a balanced design comprised of 2 classes of sites having equal sizes; hence, the prior probability of each site class is equal. Four aspects of codon evolution ($\omega$, c, $\kappa$, and $\pi$) are permitted to be heterogeneous among site classes. As we are interested in detecting sites subject to positive selection, site classes were always heterogeneous for $\omega$ ($\omega_1 = 0.3$ and $\omega_2 = 1.5$). The study (table 1) is designed to cover all $2^3 = 8$ possible combinations of heterogeneity for $\kappa$'s (scenarios A and C: $\kappa_1 = 1$, $\kappa_2 = 5$), branch lengths (scenarios A and B: $c_1 = 1$, $c_2 = 5$), and codon frequencies (in A1, B1, C1, and D1: $\pi_{1i} = 1/64$ and $\pi_{2i}$ = empirical estimates from abalone sperm lysin). Note that the generating models A1 and A2 match the models assumed under LiBaC1 and LiBaC2, respectively. Data sets of 1,000 codons are created by independently simulating 2 groups of 500 codons. Data are simulated on a 16-taxon tree (Supplementary Material online). For each scenario, 100 data sets are simulated by using the program "evolver" of the PAML package (Yang 1997). All data sets are analyzed under soft-LiBaC1, soft-LiBaC2, and hard-LiBaC2.

In addition, each scenario is analyzed under the most commonly used random effect models (M2a, M3, and M8). Here, branch lengths and parameters of the substitution model (except $\pi$'s) are obtained by ML. Placement of sites into groups is carried out by using the naive empirical Bayes (Nielsen and Yang 1998) and Bayes empirical Bayes methods (Yang et al. 2005), hereafter referred to as NEB and BEB, respectively. The generating model D2 matches the random effect model M3 ($G = 2$).

Bayes error rates (table 2) provide a measure of difficulty of each scenario. For example, scenario A2 should be the easiest and D2 the hardest. Moreover, there is a clear relationship between the level of difficulty and the specification of the $c$ parameter; scenarios where parameter $c$ differed among groups (A and B: $c_1 = 1$ and $c_2 = 5$) represent substantially easier cases as compared with scenarios where parameter $c$ is the same among partitions (C and D: $c_1 = c_2 = 1$). The difference in $\omega$ and $c$ in scenarios A and B yields a difference in the absolute

**Table 2**
**The Bayes Error Rate and Method-Specific Classification Error Rates under the 8 Scenarios with the 500–500 Site Partition of Simulation Study 1**

| | Simulation Scenarios | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A1 (%) | A2 (%) | B1 (%) | B2 (%) | C1 (%) | C2 (%) | D1 (%) | D2 (%) |
| Bayes error rate | 8.69 | 8.00 | 9.29 | 8.69 | 20.84 | 21.92 | 23.26 | 26.79 |
| Observed error rate | | | | | | | | |
| Soft-LiBaC1 | 8.98 | 10.81 | 9.75 | 9.24 | 23.66 | 23.31 | 26.13 | 28.85 |
| Soft-LiBaC2 | 9.35 | 8.59 | 10.16 | 9.09 | 23.96 | 23.62 | 29.07 | 28.29 |
| Hard-LiBaC2 | 9.52 | 8.63 | 10.25 | 10.65 | 25.78 | 26.50 | 29.59 | 29.60 |
| M3 (NEB) | 11.53 | 10.42 | 10.16 | 9.20 | 47.33 | 47.01 | 43.90 | 41.85 |
| M8 (NEB) | 11.83 | 10.97 | 10.40 | 9.53 | 49.89 | 49.73 | 49.61 | 49.36 |
| M2a (NEB) | 12.30 | 11.46 | 10.92 | 9.94 | 49.94 | 49.84 | 49.87 | 49.82 |
| M8 (BEB) | 18.81 | 17.47 | 18.35 | 17.52 | 49.89 | 49.88 | 49.79 | 49.82 |
| M2a (BEB) | 23.47 | 21.21 | 21.36 | 20.44 | 49.82 | 49.86 | 49.73 | 49.51 |

NOTE.—LiBaC1 assumes heterogeneous codon frequencies ($\pi_i$'s) among data partitions. LiBaC2 assumes homogenous codon frequencies ($\pi_i$'s) among data partitions.

rate of $d_S$ and $d_N$ between groups of sites, which likely contributes to a greater signal for differential evolution between the 2 classes of sites. For any given classifier in table 2, the observed misclassification rate corroborates the relative difficulty of a scenario inferred from the Bayes error rate.

When comparing among classifiers, LiBaC nearly achieves the theoretical lower bound on misclassification error under cases A1 and A2. This is expected, as in these scenarios, the generating model matches that employed under LiBaC1 and LiBaC2, respectively. In scenarios A2, B1, and B2, the misclassification error of some classifiers (M2a, M3, and M8 under NEB) is comparable with LiBaC, although not quite as close to the lower bound. This is noteworthy because the assumption that $c_1 = c_2 = 1$ under M2a, M3, and M8 is incorrect for A2, B1, and B2, respectively. The reason for the low misclassification rates for those models is that differences among sites in both $d_S$ and $d_N$ can be "absorbed" by the among-site variability in the ω parameter alone. We note that among-site heterogeneity in both $d_S$ and $d_N$ is a realistic aspect of gene sequence evolution as several studies have uncovered evidence of such variability in a wide variety of genes and genomes (e.g., Kosakovsky Pond and Muse 2005; Bao et al. 2007). We believe this explains, in part, why models M2a, M3, and M8 have been so successful in providing biologically valuable results for a wide variety of genes (e.g., Yang 2005). However, our simulations indicate that estimates of ω under models M2a, M3, and M8 can be biased in such cases due to misspecification for the $c$ parameter (Supplementary Material online).

In all the scenarios, the LiBaC methods are closer to the theoretical lower bound than the other classifiers (table 2). Under scenarios C and D, the observed error rates for all the classifiers are higher and further from the theoretical lower bound as compared with scenarios A and B. Interestingly, the 3 LiBaC methods outperformed M3 in case D2, where the generating model matches that of M3. The results suggest that LiBaC-based methods can be robust to model misspecification.

The misclassification error rates in table 2 are based on a posterior probability cutoff of 50%. However, the usual practice with NEB and BEB methods is to choose a higher cutoff value for sites having $\omega > 1$ (typically 90% or 95%) thereby increasing the precision of identifying sites subject to positive selection. The cost of this practice is reduced recall. In order to evaluate the classifiers in this context, we plotted the precision and recall of each classifier in relation to an idealized precision–recall curve (fig. 1). The precision–recall curve is obtained from the Bayes error rate and represents the theoretical upper bound on performance for a given scenario. In general, curves that closely follow the upper and right side borders of the parameter space (scenarios A and B) are theoretically easier classification problems than those with curves that approach the 45° diagonal (scenarios C and D). As expected, the observed relationship between precision and recall for individual classifiers falls at or below the curve (fig. 1). The spread of points for a particular cutoff posterior probability (i.e., taking results for only 0.5, 0.9, or 0.95) illustrates that using common posterior probability cutoff does not yield comparable levels of performance among different classifiers. To directly compare precision among different classifiers, we would need to fix their recall at a common level, and this is not practical. However, we can assess each classifier by its distance from the precision–recall curve.

Consistent with the notion that scenarios A and B represent the easier cases; all classifiers performed reasonably well, being close to the upper bound on performance (fig. 1). Taken across the different cutoff values, the different classifiers achieve a remarkably wide range of trade-offs between precision and recall. Consistent with the misclassification rates, all 3 LiBaC methods are closer to the precision–recall curve than the other classifiers (fig. 1 and table 3).

Scenarios C and D present a bigger challenge to all classifiers. The NEB and BEB methods under models M2 and M8 achieve acceptable levels of precision but only by a nearly complete loss of recall (fig. 1 and table 3). These classifiers cannot be improved by adjusting the posterior probability cutoff as recall remains close to zero when the cutoff is set to 50%. The trade-off between recall and precision for NEB under model M3 can be adjusted via the posterior probability cutoff; but in these scenarios, this classifier falls well below the idealized precision–recall curve (fig. 1). As LiBaC methods yield performance close

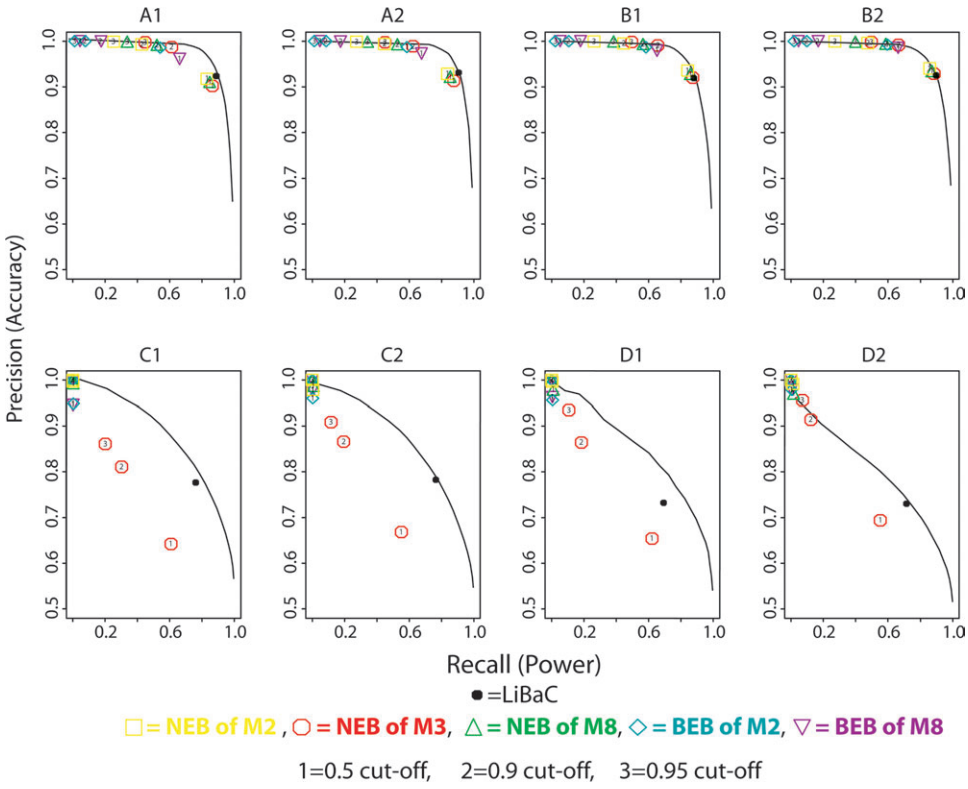


FIG. 1.—Performance of LiBaC, and selected alternative methods, relative to the theoretical upper bound on precision for the 8 simulation scenarios described in table 1. The curve illustrates the relationship between the upper bound on precision relative to recall. Note that precision and recall are equivalent to the measures called "accuracy" and "power" by Anisimova et al. (2002).

to the idealized precision–recall curve in scenarios C and D, its performance can be further tuned by adjusting the posterior probability cutoff to achieve different trade-off between precision and recall.

Measuring performance relative to the recall and precision curves in figure 1 reflects the primary interest of biologists to minimize the number of falsely detected positive selection sites among the number of sites they have inferred from the data; this is why the curves focus on precision rather than the classical false-positive rate (1—the specificity of the classifier). In cases A and B, there is a clear "elbow" (i.e., where the curvature is the greatest); this is the point where there is an optimal trade-off between precision and recall. Although the cutoff could be adjusted under LiBaC, simply using a 50% cutoff yields precision in excess of 90% while maintaining recall at 88–89% (table 3).

### *Simulation Study 2: Realistic Purifying Selection Scenario*

Although the evolutionary models permitted by LiBaC1 and LiBaC2 are quite complex, they are nonetheless highly idealized as compared with the complexity of real gene evolution. In order to evaluate performance closer to a real-world scenario, we assume that sites were drawn from an unbalanced mixture of site classes characterized by continuous distributions for ω (fig. 2A). Site class 1 is simulated under a beta distribution for ω having a mean = 0.1, with $p = 0.5$ and $q = 4.5$ (L shaped), and site class 2 sites

**Table 3**
**Precision and Recall of Positively Selected Sites in Scenarios A1, A2, and D2 of Simulation Study 1**

| | | Soft-LiBaC | | Hard-LiBaC | NEB | | | BEB | |
|---|---|---|---|---|---|---|---|---|---|
| | | Model 1 | Model 2 | Model 2 | M3 | M2 | M8 | M2 | M8 |
| A1 | Precision | 93.5 | 92.7 | 92.0 | 90.2 (99.8) | 91.8 (99.9) | 91.1 (99.9) | 98.6 (100) | 96.1 (100) |
| | Recall | 88.2 | 88.3 | 89.2 | 86.3 (44.7) | 83.0 (25.2) | 84.7 (33.7) | 54.1 (1.1) | 66.1 (4.4) |
| A2 | Precision | 91.9 | 93.0 | 92.3 | 91.3 (99.8) | 92.9 (99.9) | 92.2 (99.9) | 98.7 (100) | 97.3 (100) |
| | Recall | 86.1 | 89.5 | 90.2 | 87.5 (45.0) | 83.8 (27.4) | 85.4 (34.2) | 58.6 (1.3) | 67.6 (4.8) |
| D2 | Precision | 69.5 | 69.4 | 70.2 | 69.4 (95.6) | 99.1 (100) | 97.1 (100) | 97.8 (100) | 97.9 (100) |
| | Recall | 75.3 | 77.6 | 71.0 | 55.2 (7.0) | 1.2 (0) | 1.6 (0) | 0.4 (0) | 0.4 (0) |

NOTE.—NEB denotes naive empirical Bayes, and BEB denotes Bayes empirical Bayes. For NEB and BEB, precision and recall are shown for a posterior probability cutoff of 50% and, in parentheses, 95%. Note that precision and recall are equivalent to the measures called "accuracy" and "power" by Anisimova et al. (2002).

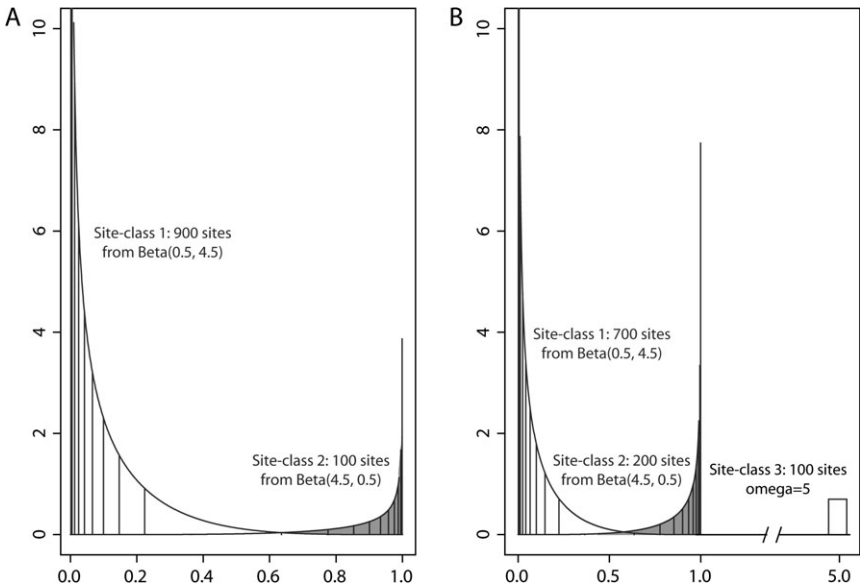


FIG. 2.—Distribution of ω specified in simulation studies 2 and 3. (*A*) Distribution of ω in a realistic scenario of purifying selection pressure (study 2). The scenario is unbalanced, being comprised of a group of 900 sites and a group of 100 sites that differ in ω and π. Group 1 (900 sites) is characterized by an L-shaped distribution for ω (beta distribution having shape parameters $p = 0.5$ and $q = 4.5$). To simulate these data, we discretize the beta distribution into 10 classes of sites having equal probability; the values of ω in each class are 0.0009, 0.0047, 0.0124, 0.0245, 0.0419, 0.0659, 0.0991, 0.1469, 0.2236, and 0.6359. Group 2 (100 sites) is characterized by a J-shaped distribution for ω (beta distribution having shape parameters $p = 4.5$ and $q = 0.5$). The area under the group 2 curve is shaded in gray. Again we use 10 discrete classes of sites; the values of ω in each class are 0.3641, 0.7764, 0.8531, 0.9009, 0.9341, 0.9581, 0.9755, 0.9876, 0.9953, and 0.9991. Discrete site classes are indicated by the vertical lines. (*B*) Distribution of ω in simulation study 3 extends simulation study 2 by adding a third group of sites having $\omega > 1$. The evolutionary models of groups 1 and 2 are the same as that in study 2, but the size is adjusted; groups 1 and 2 are comprised of 700 and 200 sites, respectively. Group 3, having $\omega > 1$, is comprised of 100 sites.

are simulated under a beta distribution for ω having a mean = 0.9, with $p = 4.5$ and $q = 0.5$ (J shaped). These groups are also heterogeneous for $c_i$ and $\pi_i$ ($\kappa_1 = 2$, $c_1 = 1$, and $\pi_1 =$ biased; $\kappa_2 = 2$, $c_2 = 5$, and $\pi_2 = 1/64$). Site class 1 is comprised of 900 sites, and site class 2 is comprised of 100 sites. Note that this scenario represents a gene that is subject only to purifying selection pressure.

In a real-world scenario, we also lack knowledge of the optimal number of site classes. Hence, prior to analysis under LiBaC, we employed the surrogate test ($\alpha = 0.01$) for number of site classes; $G = 2$ was selected for 47% of the data sets and $G = 3$ for the remaining 53% (Supplementary Material online). Given these data are comprised by a mixture of 2 continuous distributions, these results indicate that the surrogate test is likely to be conservative for real data sets. In this simulation study, we employ LiBaC under both $G = 2$ and $G = 3$.

All the LiBaC ($G = 2$) methods resolved the unbalanced structure of the data, revealing a large group of sites with low ω and a small group of sites with higher ω (table 4). For all 3 LiBaC-based methods, the parameter estimates are biased, with LiBaC having underestimated ω for both groups of sites (table 4). The bias in mean ω arises from the type of site misclassification errors; some sites from site class 1, which have generally lower ω, are incorrectly classified into site class 2 thereby causing an overestimate of its size and reducing its average ω (table 4). In this scenario, we observed a substantially larger misclassification error for hard-LiBaC, as compared with soft-LiBaC (table 4). This is presumably a result of compromising the values in the E-step under hard-LiBaC in order to limit the computational costs. Nonetheless, errors are conservative under all the LiBaC-based methods with respect to the most common analytical use of codon models; that is, in no cases was the presence of positively selected sites falsely suggested under LiBaC.

Although models M3 and M8 had the lowest overall misclassification error, this result is offset by the biases in parameter estimates that lead to false signal for sites subject to positive selection (table 4). Interestingly, M2a also produced false signal for positive selection, despite restrictions placed on its parameters intended to achieve conservative performance in detecting positive selection sites. Indeed, M2a suggests a lower average fraction of sites with $\omega > 1$, making it more conservative than M3 and M8; however, the estimated value of ω for site class 2 is largest under M2a. These findings illustrate that differences among sites in the estimated value of ω can reflect more than differences in selection pressure when other aspects of the substitution process are not constant across sites.

With the exception of hard-LiBaC, results under $G = 3$ were very similar to those obtained under $G = 2$ (Supplementary Material online). Because hard-LiBaC yielded false signal for positive selection under $G = 3$, we do not recommend it for real data analysis.

LiBaC is implemented to address the tasks of parameter estimation and site classification. For the random effect models M2a and M8, the task of classification (alternatively referred to as a "prediction" problem) has been intrinsically linked to the process of hypothesis testing. Indeed, interpretation of NEB- and BEB-based classification of positively selected sites is recommended only if an LRT for positive selection indicates that such sites exist in a given data set

**Table 4**
**Error Rate and Parameter Estimates for LiBaC and Selected Alternative Methods under the Realistic Purifying Selection Scenario (simulation study 2: 2-class beta distribution scenario with 900–100 site partition)**

| | | Average Parameter Estimates | |
|---|---|---|---|
| | Classification Error | Site Class 1 | Site Class 2 |
| Soft-LiBaC1 | 10.06% | $\omega_1 = 0.04$ (0.01)<br>$[c_1 = 1]$<br>$\kappa_1 = 2.11$<br>$p_1 = 0.79$ | $\omega_2 = 0.72$ (0.07)<br>$c_2 = 4.33$<br>$\kappa_2 = 1.98$<br>$[p_2 = 0.21]$ |
| Soft-LiBaC2 | 9.66% | $\omega_1 = 0.05$ (0.02)<br>$[c_1 = 1]$<br>$\kappa_1 = 2.11$<br>$p_1 = 0.80$ | $\omega_2 = 0.72$ (0.07)<br>$c_2 = 4.60$<br>$\kappa_2 = 1.98$<br>$[p_2 = 0.20]$ |
| Hard-LiBaC2 | 16.57% | $\omega_1 = 0.002$ (0.006)<br>$[c_1 = 1]$<br>$\kappa_1 = 2.12$<br>$[p_1 = 0.71]$ | $\omega_2 = 0.76$ (0.05)<br>$c_2 = 4.45$<br>$\kappa_2 = 1.97$<br>$[p_2 = 0.29]$ |
| M2 | N.A. | $\omega_0 = 0.06$ (0.01); $[\omega_1 = 1]$<br>$[c_1 = 1]$; $[c_1 = 1]$<br>$\kappa_0 = 2.04$; $\kappa_1 = \kappa_0$<br>$p_0 = 0.83$; $p_1 = 0.13$ | $\boldsymbol{\omega_2 = 2.48}$(0.43)<br>$[c_1 = 1]$<br>$\kappa_2 = \kappa_0$<br>$[p_2 = 0.04]$ |
| M3 | 6.37% | $\omega_1 = 0.07$ (0.01)<br>$[c_1 = 1]$<br>$\kappa_1 = 2.03$<br>$p_1 = 0.86$ | $\boldsymbol{\omega_2 = 1.45}$ (0.13)<br>$[c_1 = 1]$<br>$\kappa_2 = \kappa_1$<br>$[p_2 = 0.14]$ |
| M8 | 5.55% (5.95%) | $p = 0.33$; $q = 2.51$<br>$[c_1 = 1]$<br>$\kappa_1 = 2.04$<br>$p_1 = 0.91$ | $\boldsymbol{\omega_2 = 1.92}$ (0.29)<br>$[c_1 = 1]$<br>$\kappa_2 = \kappa_1$<br>$[p_2 = 0.09]$ |

NOTE.—Parameter values are averages of the ML estimates from 100 simulated data sets except for those shown in square brackets; values in brackets are a consequence of the clustering in the case of hard-LiBaC or are not free model parameters in other cases. The standard deviation of ω is given within parentheses. The inferred frequencies of site classes 1 and 2 are given by $p_1$ and $p_2$. Error rates for M3 and M8 are for NEB classification of sites. Error rates in parentheses for M8 are for BEB classification of sites. Classification error rates were not computed for M2 because it requires three sites classes. In no case was the true value of ω greater than 1; cases of incorrect estimation of ω greater than 1 are shown in bold.

(e.g., Wong et al. 2004; Yang et al. 2005). We did not carry out such LRTs in simulation study 1 as those simulated data always contained sites subject to positive selection. In simulation study 2, we must also consider the false-positive rate of the involved LRTs; that is, the percentage of false rejections of the null hypothesis as formulated under models M1a and M7. Hence, we carried out the required LRTs (M1a against M2a and M7 against M8) and found that the null hypothesis of no sites having evolved under positive selection was strongly rejected in all cases (Supplementary Material online). Furthermore, in all data sets, there are cases of high posterior probabilities for sites subject to positive selection under NEB and BEB, with some sites having a posterior probability >90% (Supplementary Material online). We note that several authors have predicted that model violations could lead to increases in the false-positive rate over the nominal level of the LRT and encouraged the report of such findings (Wong et al. 2004; Yang et al. 2005). Simulation study 2 clearly represents such a scenario.

### *Simulation Study 3: Realistic Positive Selection Scenario*

Simulation study 2 was based on a realistic model of a gene subject only to purifying selection pressure. In simulation study 3, we extend the model to include a small fraction of sites subject to positive Darwinian selection pressure. Here the model of evolution assumes 3 classes of sites (fig. 2B). Site classes 1 and 2 follow the same evolutionary model as in simulation study 2 above but are comprised of just 700 and 200 sites, respectively. Site class 3 is simulated under $\omega_3 = 5$, $\kappa_3 = 2$, and $c_3 = 10$ and biased codon frequencies. The surrogate test ($\alpha = 0.01$) for number of site classes indicated $G = 3$ for 99% of the replicates (Supplementary Material online).

Misclassification error rates are generally similar among all the methods (table 5). All methods resolved a small fraction of sites (estimates ranging from 9% to 18%) having $\omega > 1$ (table 5). However, M3 also suggested a second, moderately sized group of sites (22%) having ω slightly above 1, giving a false signal for a second group of sites potentially subject to positive selection. Estimates of ω under hard-LiBaC tend to be unrealistically high; again, computational savings achieved under hard-LiBaC appear to come with a cost in performance. Under models M2a, M3, and M8, heterogeneity in several aspects of the substitution process leads to amplification of the signal for differences between groups of sites, thereby allowing reasonably good classification based on ω alone. Whereas this yields favorable results when positively selected sites truly exist in the data, the effect appears to be a potential liability when such sites are not present in the data (as in simulation study 2).

### *Simulation Study 4: Strictly Neutral Model*

We simulate 50 replications under a scenario that is generally considered among the most difficult of cases; the single class of strictly neutral evolution ($\omega = 1$) scenario (Anisimova et al. 2001; Kosakovsky Pond and Frost

**Table 5**
**Error Rate and Parameter Estimates for LiBaC and Selected Alternative Methods under the Realistic Positive Selection Scenario (simulation study 3: 3-class beta distribution scenario with 700–200–100 site partition)**

| | | Average Parameter Estimates | | |
|---|---|---|---|---|
| | Classification Error | Site Class 1 | Site Class 2 | Site Class 3 |
| Soft-LiBaC1 | 15.66% | $\omega_0 = 0.05$ (0.02) | $\omega_1 = 0.57$ (0.12) | $\boldsymbol{\omega_2 = 2.87}$ (0.79) |
| | | $[c_0 = 1]$ | $c_1 = 4.28$ | $c_2 = 10.53$ |
| | | $\kappa_0 = 2.08$ | $\kappa_1 = 2.04$ | $\kappa_2 = 1.98$ |
| | | $[p_0 = 0.61]$ | $[p_1 = 0.22]$ | $[p_2 = 0.17]$ |
| Soft-LiBaC2 | 16.43% | $\omega_0 = 0.05$ (0.02) | $\omega_1 = 0.57$ (0.12) | $\boldsymbol{\omega_2 = 2.91}$ (0.84) |
| | | $[c_0 = 1]$ | $c_1 = 4.44$ | $c_2 = 10.90$ |
| | | $\kappa_0 = 2.09$ | $\kappa_1 = 2.02$ | $\kappa_2 = 2.00$ |
| | | $[p_0 = 0.62]$ | $[p_1 = 0.21]$ | $[p_2 = 0.17]$ |
| Hard-LiBaC2 | 15.28% | $\omega_0 = 0.06$ (0.02) | $\omega_1 = 0.89$ (0.18) | $\boldsymbol{\omega_2 = 74.68}$ (40) |
| | | $[c_0 = 1]$ | $c_1 = 7.03$ | $c_2 = 8.3$ |
| | | $\kappa_0 = 2.07$ | $\kappa_1 = 2.04$ | $\kappa_2 = 1.92$ |
| | | $[p_0 = 0.67]$ | $[p_1 = 0.24]$ | $[p_2 = 0.09]$ |
| M2 | 15.1% (14.6%) | $\omega_0 = 0.05$ (0.01) | $[\omega_1 = 1]$ | $\boldsymbol{\omega_2 = 3.29}$ (0.18) |
| | | $[c_0 = 1]$ | $[c_1 = 1]$ | $[c_2 = 1]$ |
| | | $\kappa_0 = 2.00$ | $\kappa_1 = \kappa_0$ | $\kappa_2 = \kappa_0$ |
| | | $p_0 = 0.64$ | $p_1 = 0.20$ | $[p_2 = 0.16]$ |
| M3 | 15.88% | $\omega_0 = 0.05$ (0.01) | $\boldsymbol{\omega_1 = 1.06}$ (0.24) | $\boldsymbol{\omega_2 = 3.38}$ (0.30) |
| | | $[c_0 = 1]$ | $[c_1 = 1]$ | $[c_2 = 1]$ |
| | | $\kappa_0 = 2.01$ | $\kappa_1 = \kappa_0$ | $\kappa_2 = \kappa_0$ |
| | | $p_0 = 0.63$ | $p_1 = 0.22$ | $[p_2 = 0.15]$ |
| M8 | 18.59% (11.35%) | $p = 0.15$, $q = 0.48$ $[\omega_1 = 0.74]$ | | $\boldsymbol{\omega_2 = 3.13}$ (0.20) |
| | | $[c_1 = 1]$ | | $[c_2 = 1]$ |
| | | $\kappa_1 = 2.00$ | | $\kappa_2 = \kappa_1$ |
| | | $p_1 = 0.82$ | | $[p_2 = 0.18]$ |

NOTE.—Parameter values are averages of the ML estimates from 100 simulated data sets except for those shown in square brackets; values in brackets are a consequence of the clustering in the case of hard-LiBaC or are not free parameters in other cases. The standard deviation of ω is given within parentheses. The inferred frequencies of site classes 1, 2, and 3 are given by $p_0$, $p_1$, and $p_2$. Error rates for M2, M3, and M8 are for NEB classification of sites. Error rates in parentheses for M2 and M8 are for BEB classification of sites.

2005). Because LiBaC is a clustering algorithm, we add a second class of sites under perfect purifying selection ($\omega = 0$) to the scenario. This 2-class scenario was balanced (500 sites with $\omega_1 = 0$ and 500 sites with $\omega_2 = 1$). The site classes also differ according to parameters $c$ and $\pi$ ($\kappa_1 = 2$, $c_1 = 1$, and $\pi_1 = 1/64$; $\kappa_2 = 2$, $c_2 = 5$, and $\pi_2 =$ biased). Although such an extreme distribution of ω is unlikely to be observed in a real biological data, a very similar scenario was shown to constitute a challenging case (Anisimova et al. 2001); it thereby serves as a useful benchmark.

Prior to application of LiBaC, we carried out the surrogate test ($\alpha = 0.01$) for number of site classes and selected $G = 2$ in 100% of the replications (Supplementary Material online). Although this scenario did not represent a difficult classification problem (Supplementary Material online), having ω on the boundary for positive selection ($\omega = 1$) represents a challenge for parameter estimation. The danger is that an estimate of $\omega > 1$ can easily be obtained for a group of sites, thereby yielding false signal for positive selection. Comparison of parameter estimates under LiBaC and M2a (the most conservative of the M-series models; Anisimova et al. 2001; Wong et al. 2004) reveals similar performance, with both approaches having a small propensity for estimating $\omega > 1$ (Supplementary Material online). Indeed, the maximum estimate of ω over all the replications was just 1.18 under soft-LiBaC1 and 1.84 under M2a. In accordance with previous suggestions (Anisimova et al. 2001; Wong et al. 2004; Kosakovsky Pond and Frost 2005), we recommend caution in attributing the evolution of sites to positive selection when the estimates of ω are only marginally larger than 1, even when an LRT such as in M1a versus M2a is significant (Supplementary Material online).

### *Real Data Analysis: Transmembrane Proteins*

Based on the results of our simulation studies, we recommend the following procedure for real data analysis.

Step1: Use a surrogate test to determine the number of site classes. Here we use the LRT under M3 ($\alpha = 0.01$) as the auxiliary criterion.

Step 2: Use empirical Bayes under M3, with $G$ determined from step 1, to provisionally place sites into groups for the purpose of model selection. We note that if structural information is available for the protein product of the gene, such information could be used in place of steps 1 and 2 as the basis a provisional assignment of sites to groups.

Step 3: Carry out model selection by using FE models following the suggestions of Bao et al. (2007).

Step 4: If a model is identified in step 3 that includes variable π's among groups, then use soft-LiBaC1 to group sites according to similarities in the process of evolution. For other models, use an appropriate mixture model as implemented in either PAML (Yang 1997) or HyPhy (Kosakovsky Pond et al. 2005). Hard-LiBaC is not recommended for real data analysis.

**Table 6**
**Selected Estimates of the Strength and Direction of Natural Selection in 8 Transmembrane Proteins**

| Gene | Protein Product | $N_C$ | M2 | M8 | Soft-LiBaC1 | Soft-LiBaC2 |
|---|---|---|---|---|---|---|
| *TrbL-VirB6_3* | VirB6_3 plasmid Conjugative transfer protein | 938 | **$\omega_3 = 5.8$** [$p_3 = 0.04$] ($P < 0.0001$) | **$\omega_2 = 4.30$** [$p_2 = 0.05$] ($P << 0.0001$) | $\omega_3 = 0.43$ [$p_3 = 0.11$] | $\omega_3 = 0.42$ [$p_3 = 0.09$] |
| *RfaL* | Putative lipid A core-O-antigen ligase | 403 | **$\omega_3 = 4.29$** [$p_3 = 0.06$] ($P = 0.0004$) | **$\omega_2 = 3.35$** [$p_2 = 0.10$] ($P < 0.0001$) | **$\omega_3 = 1.73$** [$p_3 = 0.12$] | **$\omega_3 = 1.31$** [$p_3 = 0.16$] |
| *ccmF* | Cytochrome c–type biogenesis protein | 635 | **$\omega_3 = 15.5$** [$p_3 = 0.01$] ($P = 0.0003$) | **$\omega_2 = 5.57$** [$p_2 = 0.03$] ($P < 0.0001$) | **$\omega_3 = 1.07$** [$p_3 = 0.05$] | **$\omega_3 = 2.80$** [$p_3 = 0.05$] |
| *nuoL3* | NADH dehydrogenase I chain N | 499 | **$\omega_3 = 12.53$** [$p_3 = 0.04$] ($P < 0.0001$) | **$\omega_2 = 10.37$** [$p_2 = 0.04$] ($P < 0.0001$) | **$\omega_3 = 1.45$** [$p_3 = 0.13$] | **$\omega_3 = 1.39$** [$p_3 = 0.16$] |
| *TrbL-VirB6_2* | VirB6_2 plasmid Conjugative transfer protein | 657 | **$\omega_3 = 32.8$** [$p_3 = 0.01$] ($P = 0.74$) | **$\omega_2 = 1.79$** [$p_2 = 0.03$] ($P = 0.02$) | $\omega_3 = 0.44$ [$p_3 = 0.17$] | $\omega_3 = 0.45$ [$p_3 = 0.19$] |
| *perM* | Putative permease perM homolog | 351 | **$\omega_3 = 2.57$** [$p_3 = 0.01$] ($P = 0.64$) | **$\omega_2 = 2.91$** [$p_2 = 0.02$] ($P = 0.02$) | $\omega_2 = 0.26$ [$p_2 = 0.46$] | $\omega_2 = 0.10$ [$p_2 = 0.69$] |
| *mivN* | Integral membrane protein (putative virulence factor) | 504 | **$\omega_3 = 5.95$** [$p_3 = 0.01$] ($P = 0.59$) | **$\omega_2 = 2.52$** [$p_2 = 0.01$] ($P = 0.04$) | $\omega_2 = 0.15$ [$p_2 = 0.32$] | $\omega_2 = 0.18$ [$p_2 = 0.21$] |
| *pgpA* | Phosphatidylglycerophosphatase A | 198 | **$\omega_3 = 35$** [$p_3 = 0.005$] ($P = 0.20$) | **$\omega_2 = 3.60$** [$p_2 = 0.03$] ($P = 0.05$) | $\omega_2 = 0.57$ [$p_2 = 0.23$] | $\omega_2 = 0.31$ [$p_2 = 0.21$] |

NOTE.— $N_C$ is the length of the gene in number of codons. Parameter values for $\omega_i$ are ML estimates for site class $i$; values for $p_i$ in square brackets are the frequency of sites. Estimates of $\omega$ consistent with positive selection are shown in bold. $P$ values in parentheses are for LRTs of M1a versus M2a and M7 versus M8. Complete results are presented in the Supplementary Material online.

Here, we analyze a set of 8 genes encoding transmembrane proteins according to the above steps. These genes were previously identified in a genomic scan of *Rickettsia* as having one or more significant LRTs for positive selection (KA Dunn, unpublished data). We restrict our sample of genes to transmembrane proteins because they are composed of hydrophobic membrane-spanning helices and hydrophilic loops that extend either into the cytoplasm or outside of the cell. Hence, these genes are expected to be composed of groups of sites having different equilibrium codon frequencies and evolutionary pressures. Indeed, step 3 of the analysis indicates 7 of the genes are best described by a model that permits different codon frequencies among groups (FE1 or FE9: Supplementary Material online). In the one case where we select a model that is not heterogeneous for codon frequencies (FE10), the gene (*pgpA*: 198 codons) is much smaller than the others (351–938 codons), and model selection could have been limited by the power of the involved LRTs. For the other 7 genes, either FE1 or FE9 was the best-fit model. FE1 and FE9 differ only in their treatment of the $\kappa$ parameter; FE1 permits heterogeneous $\kappa$ among sites, whereas FE9 specifies homogenous $\kappa$.

Because FE1 and FE9 specify variable $\pi$'s among groups of site, soft-LiBaC1 is most appropriate for step 4. For comparison, models M2a, M8, and soft-LiBaC2 are also employed in step 4. Selected results are presented in table 6 (see Supplementary Material online for full results). There is a clear discrepancy between the M-series models and LiBaC in the estimates of $\omega$ (table 6). Parameter estimates under M2a and M8 suggest the presence of positively selected sites in all 8 genes. Parameter estimates under soft-LiBaC1 and soft-LiBaC2 were generally consistent with each other but, unlike M2a and M8, suggest the presence of positively selected sites in only 3 genes (*RfaL*, *ccmF*, and *nuoL3*). In *ccmF*, the estimated value of $\omega$ is only marginally larger than 1 under soft-LiBaC-1, suggesting that the evidence for positive selection is weak for this gene. Based on the results of simulation study 2, a possible explanation for the discrepancy within these real data sets is the estimation bias that can arise when among-site heterogeneity is not accommodated (see simulation study 2).

The discrepancy between M2a and LiBaC is less dramatic if we require a significant LRT prior to interpreting the parameter estimates. Here, model M1a is employed for the purpose of conducting LRTs for the presence of positively selected sites. The LRT of M1a versus M2a is not significant in 4 of the 5 genes where LiBaC does not indicate positive selection (table 6); this is despite the estimate of $\omega > 1$ for a group of sites under M2a. In contrast, the LRT of M7 against M8 is significant for every gene in the data set (table 6). This result reinforces earlier findings (Anisimova et al. 2001; Wong et al. 2004) that the M1a–M2a LRT tends to be more conservative than other LRTs. For this reason, we suggest that it should be preferred over the M7–M8 LRT for real data analyses.

Taking together the results of our simulation studies and real data analyses, we recommend that data sets be tested for heterogeneity in aspects other than $\omega$ before testing for positive selection. Given evidence for such heterogeneity, we suggest claims for positive selection should be limited to those genes where results are consistent over multiple approaches. Although this idea of robustness is not a new one (e.g., Anisimova et al. 2001; Wong et al. 2004; Kosakovsky Pond and Frost 2005), our results

highlight that the LRT of M7–M8, and parameter estimates under M2a and M8, should be viewed with additional caution when several aspects of the substitution process are known to vary among sites. For certain classes of protein-coding genes, such as the transmembrane proteins examined here, their biology suggests that such heterogeneity should be expected. In these cases, soft-LiBaC1 offers a new framework, when applied collectively with other methods, to help avoid false positives and identify robust evidence for the action of positive selection.

## Discussion

A family of methods, called model-based clustering (MBC: Banfield and Raftery 1993; Fraley and Raftery 1998), were developed for a purpose similar to those pursued here. In MBC, the data are viewed as coming from a mixture of underlying multivariate normal (Gaussian) probability distributions, each representing a different cluster of data. Clusters are resolved by maximizing the likelihood function:

$$L(\theta_1, \theta_2, \theta_3; p_1, p_2, \ldots, p_k | x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} \sum_{k=1}^{G} p_k f(x_n | \theta_k),$$

where $p_k$ is the probability that an observation belongs to the $k$th cluster, $p_k > 0$; $\sum_{k=1}^{G} p_k = 1$. Our task differed from the one addressed by MBC; we wished to maximize the likelihood that a set of codon sites belong to a cluster under a given model of evolution; in this case, one of several codon models described in Bao et al. (2007). To achieve this we developed a new method, called LiBaC, which is an extension of the basic idea of MBC. The LiBaC algorithm can be viewed as a generalization of the MBC approach (Fraley and Raftery 1998) to Markov models of codon evolution (Goldman and Yang 1994; Muse and Gaut 1994). As compared with previous methods based on fixed-effect models (e.g., Yang and Swanson 2002), random effect models (e.g., Nielsen and Yang 1998), and counting methods (e.g., Suzuki and Gojobori 1999), LiBaC represents a novel approach to the problem of inferring amino acid sites subject to positive or negative selection pressure.

Site classifications under the LiBaC methods presented in this study were based on a 50% posterior probability cutoff value. Although we compared LiBaC with a selected set of commonly used methods, each using a typical posterior probability cutoff value from the literature, such comparisons are not straightforward. Our simulations clearly show that fixing different methods to a common posterior probability cutoff value does not guarantee comparable performance as recall can vary widely despite a common cutoff value. Furthermore, different cases of molecular evolution represent different levels of classification difficulty, with performance depending on the data at hand. As with the more commonly used methods, LiBaC performance could be fine-tuned by adjusting the posterior probability cutoff value; the question is how to adjust for optimal performance. We suggest that a priori data partitions, such as ones derived from the structure of the protein product, might be used for model selection and parameter estimation (e.g., Bao et al. 2007). The results of such an analysis could be used to simulate data from which an optimal precision–recall curve can be estimated and series of posterior probability cutoff values evaluated. Selecting a cutoff value closest to the point in the curve where the curvature is greatest, that is, the elbow should achieve a better trade-off between precision and recall for the data at hand than simply adopting the most commonly used cutoff value in the literature.

Recent work has revealed that among-site variability in synonymous substitution rates can occur in a variety of different genes and organisms, and can impact estimation of among-site variability in the $d_N/d_S$ ratio (Kosakovsky Pond and Muse 2005). Moreover, Kosakovsky Pond and Muse (2005) showed classification of positively selected sites under methods that assume a constant rate of synonymous substitution among sites can be negatively impacted. By using bivariate distributions to model among-site variability in both $d_N$ and $d_S$, Kosakovsky Pond and Muse (2005) achieved significant gains in the fit of a codon model to most of the data sets tested, although other aspects of the substitution process were treated as homogenous among sites. By using codon models with fixed effects determined from protein structure, Bao et al. (2007) showed that in addition to $d_N$ and $d_S$, codon frequencies, and $\kappa$ can differ significantly among groups of codons within a gene. Building on those findings, we carried out a series of simulation studies where groups of sites differed in several aspects of the substitution process. Our results clearly show that differences among sites in the estimated values of $\omega$ can reflect more than just natural selection pressure when several aspects of the substitution process differ. Simulation study 2 shows that it is possible for such estimation biases to lead to false signal for positively selected sites, and our analysis of a set of transmembrane proteins suggests such problems could be frequent among certain types of genes. Nonetheless, a large-scale survey of real gene sequences will be required to better gauge the risk of such errors to real data analysis.

We proposed a procedure for real data analysis that recommends using soft-LiBaC1 if model selection in step 3 yields a model having variable codon frequencies among groups. In step 3, models can be tested very rapidly as the group membership is treated as a fixed effect. An alternative approach would be to evaluate a wide variety of random effect mixture models that can be implemented in the program HyPhy (Kosakovsky Pond et al. 2005). This approach offers the advantage that likelihood scores can be directly compared with the commonly used M-series models (M1, M2a, etc.) as well as possibly providing better starting conditions from which to run the EM in LiBaC. Possible drawbacks of this alternative include a somewhat higher computational cost and the inability to permit codon frequencies to vary among sites. Regardless of how LiBaC is ultimately integrated into a real data analysis, we expect LiBaC will be most useful when implemented in conjunction with other approaches. In this way, it contributes to a "toolbox" of methods best applied to real data with the goal of identifying robust evidence for positive selection.

LiBaC can easily be extended to nucleotide or amino acid models; however, it appears most promising for codon

and amino acid models. Amino acids sites in the folded protein are subject to different microenvironments and perform different functions; hence, a better understanding of the connections between genotype and protein phenotype may be served by improving methods for grouping sites according to similarities in the underlying evolutionary process. Indeed, protein models are substantially improved by employing context-dependent substitution matrices in cases where the protein secondary structure can be treated as a fixed effect (e.g., Koshi and Goldstein 1995; Goldman et al. 1998; Robinson et al. 2003). For both codon and amino acid models, LiBaC could permit greater complexity in the amino acid replacement process between groups of sites by coupling it to models permitting adjustable exchangeabilities between amino acids (e.g., Dimmic et al. 2000; Yang 2000). Such an approach might improve both parameter estimates (e.g., ω) and classification of sites according to similarity of the evolutionary process. Lastly, it should also be possible to develop LiBaC for use on multigene data. Here the problem is that different genes in a genome might be best treated as having evolved under different models, but one does not want to specify an independent model for each gene in a very large data set. LiBaC-based methods could be used to cluster genes (rather than individual sites) into groups according to similarities in the underlying process of evolution.

## Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

We thank Edward Susko for helpful discussions and S. L. Kosakovsky Pond for comments that substantially improved this manuscript. This research was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (DG298394 to J.P.B. and DG40156 to H.G.) and a grant from Genome Canada.

## Literature Cited

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

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.

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

Banfield JD, Raftery AE. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics. 49:803–821.

Bao L, Gu H, Dunn KA, Bielawski JP. 2007. Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. BMC Evol Biol. 7(Suppl 1):S5.

Bielawski JP, Yang Z. 2004. Likelihood methods for detecting adaptive evolution. Statistical methods in molecular evolution. New York: Springer.

Celeux G, Govaert G. 1992. A classification EM algorithm for clustering and two stochastic versions. Comput Stat Data Anal. 14:315–332.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood for incomplete data via the EM algorithm (with discussion). J R Stat Soc B. 39:1–38.

Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput. 2000:18–29.

Dunn KA, Bielawski JP, Yang Z. 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics. 157:295–305.

Fraley C, Raftery AE. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J. 41:578–588.

Fukunaga K. 1985. The estimation of the Bayes error by the k-nearest neighbor approach. In: Kanal LN, Rosenfeld A, editors. Progress in pattern recognition. Vol. 2. Amsterdam (the Netherlands): North-Holland. p. 169–187.

Genton GG, Ronchetti E. 2003. Robust indirect inference. J Am Stat Assoc. 98:67–76.

Gillespie JH. 1991. The causes of molecular evolution. Oxford: Oxford Univesity Press.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

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

Gouriéroux C, Monfort A, Renault AE. 1993. Indirect inference. J Appl Econom. 8:S85–S118.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22:1208–1222.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676–679.

Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation in synonymous substitution rates. Mol Biol Evol. 22:2375–2385.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. Protein Eng. 8:641–645.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: LeCam LM, Neyman J, editors. Proceedings of the 5th Berkeley Symposium on mathematical statistics and probability. Berkeley (CA): University of California Press. p. 281–297.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics. 169:1753–1762.

McLachlan GJ, Krishnan T. 1997. The EM algorithm and extensions. New York: Wiley.

Muse SV, Gaut BS. 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, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20:1692–1704.

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

Tumer K, Ghosh J. 2003. Bayes error rate estimation using classifier ensembles. Int J Smart Eng Syst Des. 5:95–110.

Wong WS, Yang Z, Goldman N, Nielsen R. 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.
Yang Z. 2000. Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites. Pac Symp Comput Biol. 2000:81–92.
Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci USA. 102:3179–3180.
Yang Z, Nielsen R. 2000. Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.
Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 19:49–57.
Yang Z, Wong SW, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

Jianzhi Zhang, Associate Editor

Accepted May 30, 2008