Smoothed Bootstrap Aggregation for Assessing Selection Pressure at Amino Acid Sites

Joseph Mingrone,*[,1,2] Edward Susko,[1,2] and Joseph Bielawski[1,2,3]

[1]Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada

[2]Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS, Canada

[3]Department of Biology, Dalhousie University, Halifax, NS, Canada

***Corresponding author:** E-mail: jrm@ftfl.ca.

Associate Editor: Jeffrey Thorne

Abstract

To detect positive selection at individual amino acid sites, most methods use an empirical Bayes approach. After parameters of a Markov process of codon evolution are estimated via maximum likelihood, they are passed to Bayes formula to compute the posterior probability that a site evolved under positive selection. A difficulty with this approach is that parameter estimates with large errors can negatively impact Bayesian classification. By assigning priors to some parameters, Bayes Empirical Bayes (BEB) mitigates this problem. However, as implemented, it imposes uniform priors, which causes it to be overly conservative in some cases. When standard regularity conditions are not met and parameter estimates are unstable, inference, even under BEB, can be negatively impacted. We present an alternative to BEB called smoothed bootstrap aggregation (SBA), which bootstraps site patterns from an alignment of protein coding DNA sequences to accommodate the uncertainty in the parameter estimates. We show that deriving the correction for parameter uncertainty from the data in hand, in combination with kernel smoothing techniques, improves site specific inference of positive selection. We compare BEB to SBA by simulation and real data analysis. Simulation results show that SBA balances accuracy and power at least as well as BEB, and when parameter estimates are unstable, the performance gap between BEB and SBA can widen in favor of SBA. SBA is applicable to a wide variety of other inference problems in molecular evolution.

Key words: **adaptive evolution, positive selection, codon models, bootstrap, kernel smoothing, Bayes empirical Bayes.**

Article

Introduction

Identifying positively selected amino acid sites is a challenging statistical task that is important for investigating the functional consequences of molecular change (Yang 2005). Several approaches have been developed to detect positive selection within a protein [reviewed in Pond and Frost (2005) and Anisimova and Kosiol (2009)], but their reliability varies according to the properties of the data in hand. The most widely used methods employ a codon model to detect an excess in the nonsynonymous substitutions relative to synonymous substitutions ($dN/dS = \omega > 1$), which is an indication of evolution by positive selection. Proteins evolving under positive selection must retain the capacity to fold into complex structural and functional domains, so the majority of amino acid substitutions will be subject to purifying selection pressure, with $\omega < 1$ (Kimura 1968). From extensive surveys of positive selection in real genes, we expect that only a small fraction of amino acid sites will be subject to adaptive change and exhibit an $\omega > 1$ (Anisimova et al. 2007; Ge et al. 2008). The sparseness of these sites makes them challenging to identify.

Two general categories of methods for detecting positively selected amino acid sites include counting and fixed-effect methods. Counting methods employ ancestral reconstruction of codon states for all internal nodes of a phylogenetic tree to obtain counts of the synonymous and nonsynonymous changes along each of its branches. The counts inferred for a given site are used to test if $\omega \neq 1$. Some counting methods use parsimony (Fitch et al. 1997; Bush et al. 1999; Suzuki and Gojobori 1999), and others likelihood (Nielsen 2002; Nielsen and Huelsenbeck 2002; Suzuki 2004; Suzuki and Nei 2004; Pond and Frost 2005) to infer the ancestral codon states. The reconstructions are often similar, but under the likelihood approach uncertainty about the inference can be summarized via the posterior probabilities of the ancestral states. Thus, the parsimony based methods must assume that these uncertainties are irrelevant to the statistical test. While this makes the approach attractive for very large datasets where reliable reconstructions can be obtained relatively quickly (Lemey et al. 2012), widespread use is hindered by a lack of power when the level of divergence is too low or by the negative impact of substitutional saturation when the level of divergence is too high (Pond and Frost 2005).

An alternative approach is to treat each site as independently relevant to the question of evolution by positive

selection, and attempt to fit an ω parameter to the data at each site. Thus, the effect of each site on the task of ω inference is fixed. Model based testing for $\omega \neq 1$ can be carried out via a standard likelihood ratio test (LR), and no assumptions are required about the distribution of selection pressure, ω. Although ω is treated as a site-specific variable, other important variables in the codon model (e.g., branch lengths) are shared among sites, with their values estimated jointly from the complete set of sites. Results obtained using these modeling ideas (Massingham and Goldman 2005; Pond and Frost 2005) are encouraging, and we expect this family of methods will continue to have a role in real data analyses (Scheffler et al. 2014). However, χ^2 approximations to the distribution of the test statistic assume relatively large numbers of taxa, which is often not the case. The lack of independence of data across taxa that is due to phylogeny creates further difficulties for χ^2 approximations.

A third approach for detecting positive selection at amino acid sites, which is the focus of this article, treats the value of ω at a site as the realized value of a random variable. A particular model for the distribution of ω is chosen and maximum likelihood (ML) is used to fit the distribution to the data as part of an explicit model of codon evolution. There are recommendations (Yang and Nielsen 1998) to use a prescreen that fits two models: one with a distribution that excludes values of $\omega > 1$, and another with the same distribution, except with weight on values of $\omega > 1$ permitted. This nested-model pre-screening is used to test if the data conveys any evidence of positive selection. When the null hypothesis of no positive selection is rejected using a LR test, site-wise analysis is warranted. Site-wise analysis is carried out using Bayes rule to calculate the posterior probability that a site h evolved under some estimated value of ω, given the data at site h. This approach is referred to as empirical Bayes (EB) because the marginal distribution of ω is determined from the data. Conclusions regarding the evolution at a site are made based on the estimated ω-values along with their associated posterior probabilities conditioned on the data at the site. For example, when the largest posterior probability for a site is associated with a value of $\omega > 1$, this is taken as evidence of positive selection at that site.

Because the marginal distribution of ω is determined from the data and the site posterior probabilities always depend on the fitted values of the model parameters (shape parameters of the distribution, edge lengths, etc.), the reliability of EB inference depends on the accuracy of the fitted values. If they have been accurately estimated, as is often the case with large, information-rich datasets, they can simply be treated as known without errors. This approach is known as the naïve empirical Bayes approach (NEB) (Nielsen and Yang 1998). However, when the fitted values are subject to large errors, the detection of positive selection according to the posterior probabilities can be negatively impacted and in some cases the false positive rate can be unacceptably high (Wong et al. 2004). Bayes empirical Bayes (BEB), has been used to adjust for uncertainty in the parameters of the ω distribution by assigning priors to those parameters and using numerical integration to average over the uncertainty represented by the priors (Yang et al. 2005). Because this tactic can substantially reduce the false positive rate relative to NEB in problematic datasets, BEB has become a popular method for inferring the action of selection at individual sites. A fully Bayesian approach that also assigns priors to edge-lengths and other parameters is available for the inference of positive selection at sites (Aris-Brosou 2003; Huelsenbeck and Dyer 2004), but it is not as widely employed as EB because it is available for a limited set of models.

BEB does have limitations. As currently implemented, the BEB approach only accommodates uncertainty in the parameters of the ω distribution, leaving all others fixed to their fitted values. Furthermore, only uniform priors are used, which means the adjustment for uncertainty is independent of the signal in the data. Although these will not be serious limitations for many analyses of real data, we show through simulation and real data analysis that deriving the adjustment for parameter uncertainty from the data can improve inference for some datasets. To avoid the need for priors, we developed a new approach that uses bootstrapping (Efron 1979, 1982) of site patterns to simulate dataset variability and adjust for the uncertainty in the data. From bootstrap datasets, the distribution of the maximum likelihood estimates (MLEs) are estimated. The posterior probabilities for positive selection at a site is then obtained using an aggregate value coming from MLEs over bootstrapped data sets, rather than according to a single posterior probability obtained under NEB or BEB. In principle, bootstrap-based methods should use as many replicates as possible to approximate the infinite-sample bootstrap distribution. As this is computationally expensive, we use smoothing techniques borrowed from kernel density estimation (Silverman and Young 1987; Davison and Hinkley 1997, Section 3.4) to obtain an approximation with less computational cost. We refer to this new approach as smoothed bootstrap aggregation (SBA). Our simulation results show that SBA balances accuracy and power at least as well as BEB.

We also investigated the behavior of ML estimation when standard regularity conditions, such as the requirement that true parameter values be in the interior of the parameter space, are not met. Codon models fit ω distributions that, for some data-generating settings, violate regularity conditions, which leads to substantial instability in parameter estimation. These instabilities have a negative impact on the inference of positive selection under EB, and we show that our new approach is an improvement over both NEB and BEB in such cases. We also show that results previously reported for the *tax* gene of HTLV (Suzuki and Nei 2004) are likely a consequence of such instabilities. The *tax* gene is a well known example where EB is widely considered unreliable, and it has been used to criticize the overall approach. We provide an explanation for the previous results obtained under EB methods for the *tax* gene, and show that SBA can help diagnose such dubious inferences.

New Approaches

We developed a new approach for classifying sites we call smoothed bootstrap aggregation (SBA), which uses bootstrapping and kernel smoothing techniques to accommodate

uncertainties in MLEs. Site patterns from a sequence alignment are sampled with replacement to create a number of bootstrap sequence alignments. For each of the bootstrap sequence alignments, MLEs are calculated. The usual bootstrap distribution is the empirical distribution of the calculated MLEs. To avoid difficulties due to (1) low information content in the data, (2) necessarily limited bootstrap sampling, and (3) instabilities in the parameter estimates, we used, instead, a kernel density estimate of the bootstrap distribution coming from the MLEs. The smoothness of the distribution is controlled by a bandwidth parameter, which we set larger than conventional values to give greater smoothing.

While typical applications of bootstrapping use MLEs to calculate confidence intervals and standard errors, we, instead, use the bootstrap to accommodate uncertainty in the posterior probabilities of positive selection at sites. For any given site in the original sequence alignment, many parameter values are generated from the smoothed bootstrap distribution and substituted into posterior probability formulas to give a distribution of posterior probabilities which reflects parameter uncertainty. The mean or median of these posteriors is a more stable estimate of the true posterior and is used for classification. See figure 1 for an overview of SBA.

Results

Non-Standard ML Estimation Behavior

Parameter estimation by ML has attractive statistical properties, including consistency, efficiency, and asymptotic normality, when certain regularity conditions hold (Kalbfleisch 1985; Bickel and Doksum 2006). For settings where regularity conditions hold, we verified that we could obtain well-behaved estimates of the parameters of the ω distribution under two commonly used codon models: M2a (Nielsen and Yang 1998; Yang et al. 2005) and M8 (Yang et al. 2000a). We simulated 100 datasets representing a *regular* estimation problem with an ω distribution having at least 10% weight on each site class

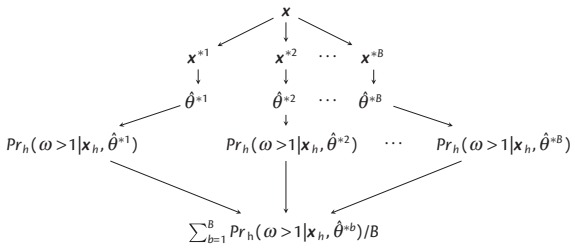

Fig. 1. Bootstrapping site patterns in a codon sequence alignment to classify selection pressure at codon sites. From an alignment of protein coding DNA sequences, $\boldsymbol{x}$, with n codon sites, site patterns are randomly sampled with replacement to obtain a bootstrap sample, $\boldsymbol{x}^{*b}$ with n sites. MLEs, $\hat{\theta}^{*b}$, are then estimated for bootstrap sample $\boldsymbol{x}^{*b}$. Using $\hat{\theta}^{*b}$ and $\boldsymbol{x}$, the posterior probability $Pr_h(\omega>1|x_h,\hat{\theta}^{*b})$, that site h is under positive selection is calculated. These steps are repeated B times to calculate B sets of posterior probabilities. An aggregate posterior probability that site h is under positive selection is calculated by, for instance, averaging posterior probabilities over bootstrap replicates, $\sum_{b=1}^{B} Pr_h(\omega>1|x_h,\hat{\theta}^{*b})/B$.

(45% $\omega=0$, 45% $\omega=0.5$, and 10% $\omega=5$). As expected, MLEs obtained from these data under both M2a and M8 have unimodal and symmetric distributions (fig. 2a and *b*). For the estimates in this *regular* case, there are no indications of departures from the limiting properties predicted by ML theory.

The regularity condition requiring true parameter values to be in the interior of the parameter space is sometimes violated when using codon models. For such parameter settings, instabilities or departures from the expected limiting properties of ML estimation can arise including non-Gaussian and over-dispersed distributions of estimates. To investigate instabilities under models M2a and M8, we simulated 100 datasets representing an *irregular* estimation problem with sparse information, i.e., 100% of the sites at the threshold for positive selection, $\omega=1$. In figure 2e and *f*, in contrast to the results presented in figure 2a and *b*, there are instabilities in the MLEs for the parameters representing the proportion of sites under positive selection, $p_{\omega>1}$. The $p_{\omega>1}$ parameter distributions under both models have mass concentrated on both the lower and upper boundaries of the parameter space, and the distributions of the corresponding $\omega_{>1}$ parameters are concentrated on the lower boundary. Application of the LR test to filter datasets that convey no evidence of positive selection did not prevent instabilities. The null hypothesis of no positive selection was rejected for 10 datasets under M2a and 9 under M8. However, the MLE distributions after applying this pre-screening step remained unstable (supplementary fig. S1, Supplementary Material online).

Some of the model M2a MLE instabilities shown in figure 2e and *f* are due to the discrete ω distribution. True discrete distributions of interest can lie on the boundary of the parameter space, which is a regularity condition violation that gives rise to MLE instabilities. For instance, consider data generated from an ω distribution with no mass on $\omega>1$. Estimates of the ω distribution will tend to approximate the true distribution and one way this can occur under M2a is when $\hat{\omega}_{>1}\approx 1$. When this happens, the likelihood will remain approximately constant over all choices of $p_{\omega=1}$ and $p_{\omega>1}$, giving a sum, $p_{\omega>1}+p_{\omega=1}$, that is approximately the same as that of the MLE. Consequently, estimates of $p_{\omega=1}+p_{\omega>1}$ are stable, but estimates of $p_{\omega=1}$ and $p_{\omega>1}$ are not, because many different choices give the same sum. Likewise, when a $p_{\omega>1}$ parameter is estimated near 0, the corresponding $\omega_{>1}$ can take on almost any value without changing the likelihood. For example, two M2a and six M8 biologically unrealistic estimates of the $\omega_{>1}$ parameter (e.g., $\omega_{>1}=999$) occurred when the corresponding $p_{\omega>1}$ parameters were estimated to be 0. These estimates were excluded from the $\omega_{>1}$ histograms. For the data representing an *irregular* estimation problem with all sites simulated with $\omega=1$, two other problematic M2a parameterizations that fit the data equally well occurred often. First, all the weight was put on the ω_1 category and second, all the weight was put on the $\omega_{>1}$ category when it was estimated very close to 1. Although there is virtually no difference in the likelihood scores between the two parameterizations, the NEB posterior probabilities for positive selection were 0 and 1 respectively. These different MLE instabilities arose with two general types

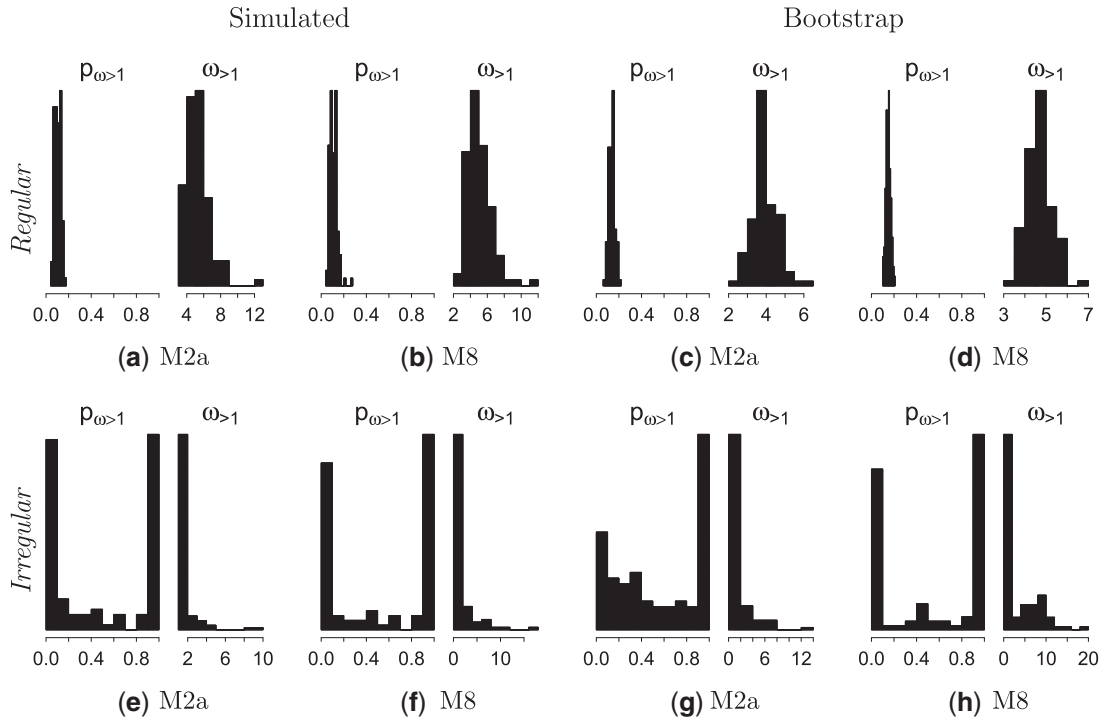

Fig. 2. MLE distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters under M2a and M8. Histograms are over 100 simulated (*a,b,e,f*) and bootstrap (*c,d,g,h*) datasets with the bootstrap datasets generated by sampling from one simulated dataset. Data were simulated under *regular* (*a–d*) and *irregular* (*e–h*) conditions. *Regular* simulation conditions: 5 taxa, 45% $\omega = 0$, 45% $\omega = 0.5$, and 10% $\omega = 5$. *Irregular* simulation conditions: 5 taxa, 100% $\omega = 1$.

of simulation settings: (1) when fewer site classes were simulated than exist in the fitted model, and (2) when different site classes were simulated with similar levels of selection pressure.

When working with real data, often only a single sample is available and alternative techniques must be used to approximate distributions of parameter estimates. One such technique is the bootstrap. We used our bootstrap-based approach with sequence alignments to investigate properties of the MLE distributions and to detect settings where inference tends to be problematic (see "Methods" section). Whereas sampling with replacement from a single sample leads to a bootstrap parameter distribution that is a jagged estimate of a smooth distribution, we found the bootstrap, in many cases, can effectively estimate the distributions of MLEs. Figure 2c and *d* shows the distribution of the ω MLEs associated with positive selection generated over 100 bootstrap samples of a *regular* dataset. Note the resemblance of the bootstrap distributions in figure 2c and *d* to the analogous distributions over simulated datasets in figure 2a and *b*. A comparison of figure 2e, *f* and 2g, *h* illustrates that when the distribution over multiple samples is problematic, so too is the distribution over bootstrap samples. Among the 100 bootstrap MLE distributions obtained from the datasets simulated under *irregular* model conditions, we identified 91 of the M2a and 95 of the M8 $p_{\omega>1}$ parameter distributions as unstable using the criterion that at least 5% mass lies both below 0.2 and above 0.8. These distributions indicate that the mixture distribution for ω "flip-flops" between few and many sites in a positive selection class. Recall that under the generating model for these data, no sites are under positive selection. Plots of the other parameters of the ω distributions can be found in supplementary figure S2, Supplementary Material online. Scenarios when the bootstrap distribution is not a good estimate of the true distribution of parameter estimates has been described in other settings (Efron and Tibshirani 1994, p. 81). Therefore, while the bootstrap alone can be helpful for identifying problems, it is not always a robust solution for deriving a correction for parameter uncertainty.

Kernel Smoothing Improves the Bootstrap-Based Method for Approximating MLE Distributions

To avoid results that are a consequence of randomness due to bootstrapping, it is beneficial to choose the number of bootstrap samples, *B*, large enough so that the finite-sample bootstrap distribution approximates the infinite-sample bootstrap distribution well. However, when regularity conditions are violated there is no guarantee that even the infinite-bootstrap distribution provides an adequate assessment of the variability of an MLE. We tested this assertion under codon models where the distributions of the $p_{\omega>1}$ parameters were unstable over simulated and bootstrap datasets. For the data representing *irregular* model conditions described above, we generated 10,000 bootstrap datasets for each of the first 10 simulated datasets. The instabilities that characterize these 10 bootstrap distributions were largely unchanged by increasing *B* (supplementary fig. S3, Supplementary Material online). Similar difficulties arise in a variety of bootstrap applications. As a simple example of the phenomenon, suppose interest is in θ from a binomial distribution with small *n* and small θ. It is possible to sample almost all zeros, in which case the variance of the bootstrap distribution of θ estimates will be too small. Such boundary issues related to small samples can similarly be problematic for ω distributions when estimated weights are close to 0.

We used kernel smoothing along with bootstrapping to characterize the uncertainty in MLEs under *difficult* estimation conditions. Kernel smoothing is typically used to approximate the infinite-sample distribution more effectively when using a smaller number of bootstrap samples. However, the standard application of this technique (Davison and Hinkley 1997, p. 79) was not sufficient when the MLEs were unstable. For such cases, *over smoothing* (i.e., using a larger than typically considered optimal bandwidth) was necessary to obtain conservative estimates of the MLE distributions, with larger variance that suppressed the influence of the instabilities (supplementary fig. S4, Supplementary Material online). By over-smoothing the p parameters of codon models M2a and M8 with a uniform kernel we compensated for (1) low information content in the data, (2) fewer bootstrap samples, and (3) instabilities in the parameter estimates. For this reason, we included over-smoothing of the p parameters in all applications of SBA.

Simulation Results

We used simulation to compare the performance of SBA with BEB and NEB. The design of our studies was motivated by the more challenging schemes of Wong et al. (2004) and Yang et al. (2005), however our design extends theirs to investigate performance under progressively more model misspecification. The design is divided into three scenarios covering three levels of model misspecification. The *Correct Model Scenario* is comprised of four simulation studies (studies 1–4) where the nuisance parameters of the generating model ($\kappa = 1$, $\pi_i = 1/61$) were freely estimated by the fitted model. The ω distributions used to generate the datasets are listed in the third column of table 1. This scenario design matches selected schemes in Yang et al. (2005). The *Mild Misspecification Scenario* uses the same ω distribution as the first scenario as the basis of four additional studies (studies 5–8), but includes mild misspecification of the nuisance parameters (see Methods). Lastly, the *Heavy Misspecification Scenario*, includes two studies (studies 9–10) with heavy misspecification for the fitted model, which represents a more plausible scenario for the analysis of real sequences. In one study (study 9), the data were simulated using the highly biased codon frequencies from the Drosophila *GstD1* gene (Bielawski and Yang 2005). In the second study (study 10), the generating model is based on a 50/50 mixture of two heterogeneous classes of sites. One class was generated using equal codon frequencies, $\kappa = 1$, and $\omega = 0.5$, while the other used the Drosophila *GstD1* gene codon frequencies, $\kappa = 8$, and $\omega = 1$. For all ten simulation studies, we simulated 100 alignments, each having 500 codons, using the same 5-taxon tree from Wong et al. (2004). The studies in the *Correct Model Scenario* were repeated under model M2a with the 30-taxon tree from the same paper.

Table 1 lists the false positive rates (proportion of sites inferred positively selected among those that are not) using a posterior probability cutoff of 0.95 for NEB, BEB, and SBA under models M2A and M8. Study 1 (no misspecification of the nuisance parameters and all sites simulated using $\omega = 1$) is an interesting case as NEB exhibits false positives, while BEB and SBA do not. This is expected; NEB is known to yield unreliable posterior probability calculations in small datasets (Anisimova et al. 2002; Yang et al. 2005). Because the conditions of study 1 yield unstable parameter estimates (fig. 2e–*h*), the false positives under NEB reflect more than mere sampling errors. MLE instabilities cause large $p_{\omega>1}$ to occur too often and these values lead to high posterior probabilities for positive selection under M2a and M8. The posterior probability calculations under SBA and BEB are reliable because those approaches do not assume the MLEs have been estimated without error. Yang et al. (2005) suggests that with more data, the problems with NEB controlling false positives can be mitigated. However, the MLE instabilities persisted in study 1 using a tree topology with 30 taxa (supplementary fig. S5, Supplementary Material online), indicating that large sample sizes do not always ensure accurate predictions.

Relative to simulations with a single $\omega = 1$ (study 1), when the ω distribution was 50% $\omega = 0.5$ and 50% $\omega = 1$ (study 2), the overall signal for positive selection was diminished and all false positive rates were 0. Conversely, when the ω distribution was 50% $\omega = 1$ and 50% $\omega = 1.5$ (study 3) there was a slight increase in the NEB false positive rates relative to study 1. Under M2a the false positive rates were 0 using BEB and SBA, but under M8 they increased to 0.05 using BEB and to 0.02 using SBA. For study 4, because the simulated ω values for the three sites classes were far enough apart, the false positive rates were well controlled.

The introduction of mild model misspecification of the nuisance parameters did not result in higher false positive rates under M2a, but did under M8. For studies 5–8, the BEB false positive rates (using a 0.95 posterior probability threshold) under M8 increased in all four cases relative to the corresponding studies (1–4) in the *Correct Model Scenario*. The same SBA false positive rates only increased in two cases and by smaller amounts than with BEB. When heavy model misspecification was introduced in the third scenario, NEB failed to adequately control false positives with rates between 50% and 71% under both M2a and M8. BEB and SBA also did not control the false positive rates in study 9, but did in study 10.

The results in table 1 are over all sites in all simulated datasets. After applying LR tests at the 0.05 level to filter datasets that convey no evidence of positive selection, none of the false positive rates under BEB or SBA changed. Supplementary table S1, Supplementary Material online gives the false positive rates under NEB after the adjustment. With the exception of two cases, the effect is minimal. Interestingly, under the null hypothesis, the false positive rates of the LR tests were larger than expected, particularly with model misspecification.

When testing for positive selection, we aim for large true positive rates, the proportion of sites truly under positive selection that are correctly identified, sometimes referred to as power. A difficulty in comparing methods for detecting positive selection is the choice of threshold. Lower thresholds tend to increase the true positive rate, but tend to also increase the false positive rate. To ensure that comparisons of power for different methods correspond to the same false positive rate we used Receiver Operator Characteristic (ROC)

Table 1. Simulation Design and False Positive Rates under NEB, BEB, and SBA each with Models M2a and M8.

Study	Misspecification	ω distribution	NEB		BEB		SBA	
			M2a	M8	M2a	M8	M2a	M8
1	None	100% 1	*0.34*	*0.35*	0.00	0.00	0.00	0.00
2	None	50% 0.5, 50% 1	0.00	0.00	0.00	0.00	0.00	0.00
3	None	50% 1 50% 1.5	*0.35*	*0.37*	0.00	*0.05*	0.00	0.02
4	None	45% 0, 45% 1, 10% 5	0.00	0.00	0.00	0.01	0.00	0.00
5	Mild	100% 1	*0.20*	*0.37*	0.00	*0.24*	0.00	*0.13*
6	Mild	50% 0.5, 50% 1	0.00	*0.13*	0.00	*0.11*	0.00	0.02
7	Mild	50% 1, 50% 1.5	*0.30*	*0.30*	0.00	*0.39*	0.00	*0.12*
8	Mild	45% 0, 45% 1, 10% 5	0.00	0.04	0.00	*0.12*	0.00	0.00
9	Heavy	100% 1	*0.71*	*0.71*	*0.55*	*0.62*	*0.13*	*0.52*
10	Heavy	50% 0.5, 50% 1	*0.53*	*0.50*	0.00	0.00	0.00	0.01

NOTE.—A posterior probability threshold of 0.95 was used for classifying sites to be under positive selection. Under SBA, smoothing was carried out using a uniform kernel with a bandwidth parameter $h = 0.4$. False positive rates 0.05 and larger are shown in italics.

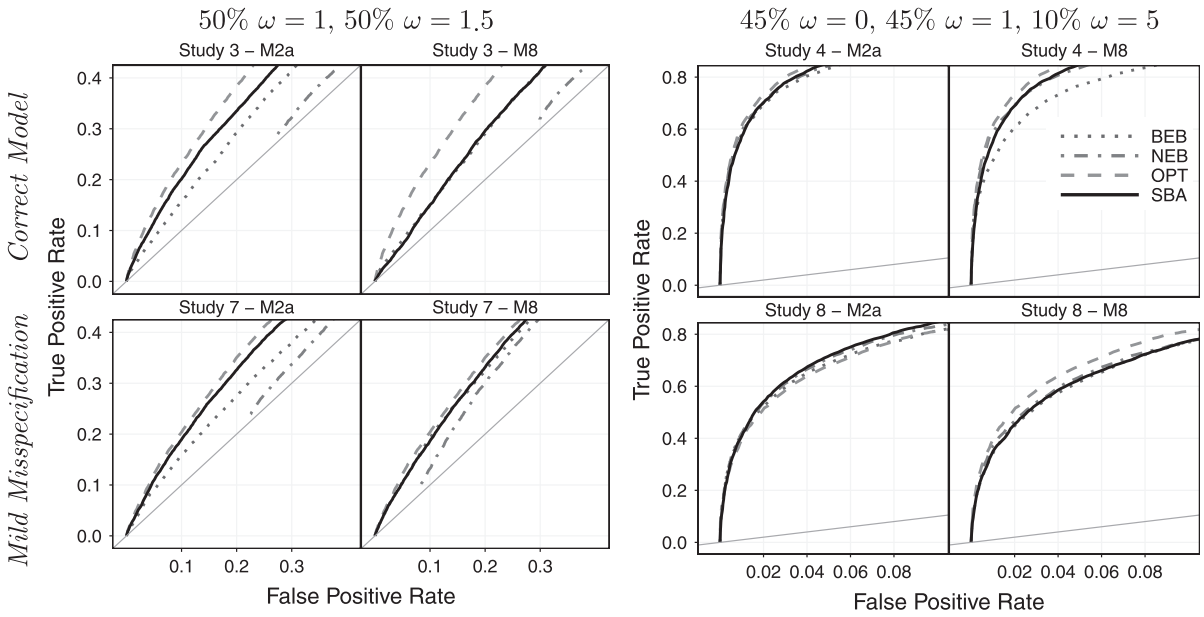

FIG. 3. ROC curves for the detection of sites under positive selection for BEB, NEB, and SBA analyses of data generated under two different simulation scenarios: without model misspecification (*Correct Model*, studies 3 and 4) and with mild model misspecification (*Mild Misspecification*, studies 7 and 8). The data were simulated using a 5-taxon tree topology. In studies 3 and 7, 50% of the sites were simulated under neutral evolution ($\omega = 1$) and 50% of the sites under positive selection ($\omega = 1.5$). In studies 4 and 8, 45% of the sites were simulated under purifying selection ($\omega = 0$), 45% under neutral evolution ($\omega = 1$) and 10% under positive selection ($\omega = 5$). Each plot includes a line for the lower bound ($y = x$) and an expected upper bound (OPT) when classification is made using the generating model parameters. Curves for NEB do not always cover the whole range of false positive rates, because NEB sometimes estimates the ω distribution with all mass on $\omega > 1$. In these cases, even with a posterior probability cut-off of 1, NEB still incorrectly classifies sites to be under positive selection.

curves, a convenient way to visualize the balance between accuracy and power for classification problems. Each point on a curve represents a threshold for the posterior probability of positive selection. Figure 3 shows ROC curves for each of the simulations that included positive selection in the generating model (studies 3, 4, 7, and 8). Curves are also included for the classification of sites using the generating parameters, i.e., the MLEs are fixed to the simulated values. These curves represent an expected upper limit in performance of site classification (supplementary file S1, Supplementary Material online). The lower limit for classification, when each site is randomly identified to be under positive selection, is represented by a $y = x$ line.

The introduction of mild misspecification made the task of detecting sites under positive selection more difficult in study 8. This is evident from the shifting of the ROC curves down and to the right (lower rates of true positives for a given false positive rate) in study 8 relative to the corresponding simulations without the misspecification of the nuisance parameters in study 4. The same effect was not observed between the ROC curves of studies 3 and 7.

In all cases, the SBA curves were at least as close as the BEB curves to the expected upper limit. In studies 3 and 7 (50% $\omega = 1$, 50% $\omega = 1.5$), under M2a, where the estimates of the $p_{\omega>1}$ and $\omega_{>1}$ parameters were unstable (supplementary fig. S6, Supplementary Material online), the gaps between

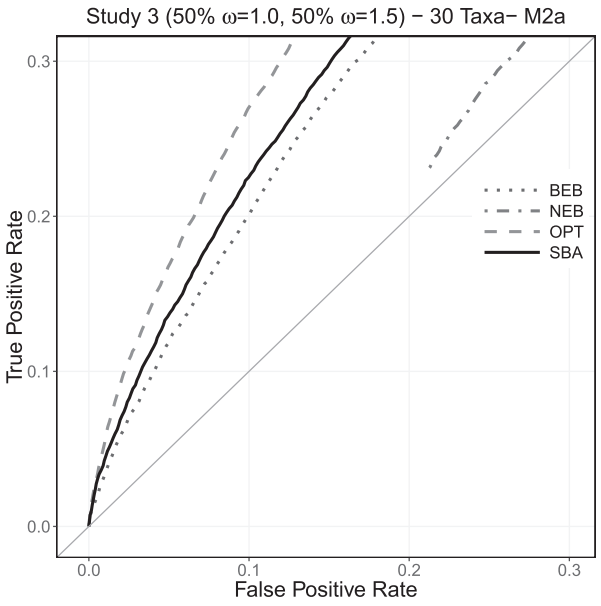

Fig. 4. ROC curves for the detection of sites under positive selection for BEB, NEB, and SBA analyses of data generated under *Correct Model*, study 3 (50% $\omega = 1$, 50% $\omega = 1.5$). The data were simulated using a 30-taxon tree topology. The plot includes a curve for the lower bound ($y = x$) and an expected upper bound (OPT) when classification is made using the generating model parameters. The curves for NEB do not always cover the whole range of false positive rates, because NEB sometimes estimates the ω distribution with all mass on $\omega > 1$. In these cases, even with a posterior probability cut-off of 1, NEB still incorrectly classifies sites to be under positive selection.

the curves for BEB and SBA were the largest, even when the number of taxa was increased from 5 to 30 (fig. 4). This indicates that SBA, for a given false positive rate, had more power to detect sites under positive selection than BEB. In studies 4 and 8 (45% $\omega = 0$, 45% $\omega = 1$, 10% $\omega = 5$), where the parameters of the ω distribution were well estimated, all approaches (NEB, BEB, and SBA) performed well and the ROC curves were all close to the expected upper limit. Taken together, the results suggest that SBA balances accuracy and power at least as well as BEB and may be preferable to BEB when parameter estimates are unstable.

Real Data Analysis

We began our analysis of the 16 real datasets (described in the "Methods" section and summarized in table 2) using the bootstrap distributions of the MLEs to investigate their properties. We examined the unsmoothed distributions of the parameters of the ω distribution. These distributions indicate that the MLEs for a given model can have very different properties in different real datasets (supplementary figs. S8–S11 , Supplementary Material online). Although the real data represent different degrees of *regular* and *irregular* model properties, we were able to identify groups of genes that represent both extremes. The *regular* cases had no clear evidence of MLE instabilities and low bootstrap variance (e.g., lysin; fig. 5a and *b*). We determined that the ω distributions had been well estimated for six genes (*pol*, *vif*, lysin, *nuoL3*, *RafL*, and *TrbL-VirB6_3*). In contrast, we uncovered evidence of MLE instabilities in other genes (e.g., *CDH3*; fig. 5c and *d*). We determined that the ω distributions had been poorly estimated for five genes (*CDH3*, *mivN*, *pgpA*, *tax*, and *TrbL-VirB6_2*) under at least one model. Because no single summary statistic (number of taxa, sequence length, tree length) was generally predictive of *irregular* model properties, we recommend visual inspection of the bootstrap distributions for all real data analyses (supplementary figs. S10 and S11, Supplementary Material online).

Next we investigated the degree to which the real data results obtained under BEB, NEB, and SBA were consistent with each other. This is challenging, because the posterior probability thresholds for site classification are not calibrated to give comparable false positive rates. Our solution was to measure the rank correlations of the site-specific posterior probability scores for positive selection between methods (BEB, NEB, and SBA). As there are a large number of pairwise comparisons, we took the mean relationship between methods for both the genes representing *regular* and *irregular* model estimation (table 3). We found that when MLEs are well estimated (*regular* genes), there is stronger agreement among all three methods in the ranking of sites according to the signal for positive selection. In contrast, when the ω distributions are poorly estimated (genes representing *irregular* estimation), BEB and SBA are generally consistent in their rankings, but differ from NEB. These results suggest that NEB's inability to accommodate MLE uncertainty in such datasets has the largest effect on the posteriors. However, the problem of calibration remains. Our simulation studies revealed that using a common posterior probability threshold for classification does not guarantee a similar trade-off between accuracy and power for different methods. Indeed, we see evidence of this in the real data. Comparing the counts of positively selected sites identified in the genes using thresholds of 0.50 and 0.95 reveals differences between BEB and SBA (table 4), despite large rank correlations. Under M2a, there was a stark difference between the *irregular* genes and all other genes. ROC curves for simulations studies are better suited for comparing methods, because they give direct comparisons of power at the same false positive rate. We also used rank correlation to investigate the robustness of the methods (BEB, NEB, and SBA) to the chosen model (M2a versus M8). We did this by computing the rank correlation, between models, of the site posterior probabilities obtained by the same method (table 5). For the *regular* genes, all three methods had high correlations with low variably. For the genes representing *irregular* estimation, the correlation was lower and the variability larger for NEB as compared with BEB and SBA. The similarity across models that we observed for SBA may be a consequence of using nonparametric bootstrapping, which should show robustness to model misspecification. It seems that BEB's application of uniform priors to the ω distribution achieved a similar effect.

Up to this point, bootstrapping has been used to obtain surrogates for posteriors. An alternative use of bootstrapping is to construct confidence intervals for posteriors to quantify the uncertainty at any given site about what the true

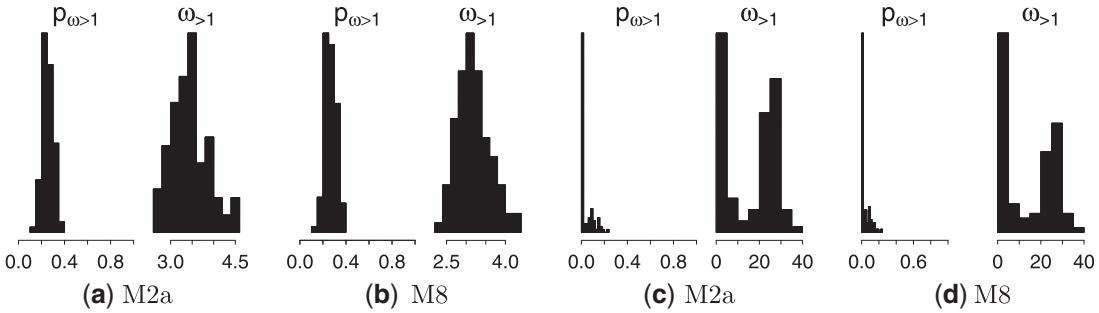

FIG. 5. MLE distributions over bootstrap datasets for the lysin and *CDH3* genes. The distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters associated with positive selection were estimated under models M2a and M8 for each of 100 bootstrap datasets.

Table 2. Genes Analyzed under Models M2a and M8 Using NEB, BEB, and SBA Approaches for Site Classification.

			−lnL				
Gene	N_t	N_c	M1a/M2a	M7/M8	*P*-value	TTL	N_s
β-globin	17	144	3716.14/3712.55	3697.22/3686.13	0.0275/1.53e−5	8.40/8.57	0(0)/3(4)
ccmF	5	635	6121.78/6113.57	6127.62/6116.48	2.72e−4/1.46e−5	5.60/3.03	3(2)/3(5)
CDH3	11	828	5629.97/5623.37	5630.66/5623.88	1.35e−3/1.14e−3	0.56/0.56	1(1)/1(1)
ENAM	11	1142	7514.30/7509.28	7609.16/7605.74	6.61e−3/0.0327	0.46/0.56	1(1)/2(1)
env	13	91	1114.64/1106.45	1115.40/1106.39	2.76e−4/1.23e−4	2.04/2.04	2(2)/2(4)
lysin	25	134	4472.65/4410.28	4472.16/4410.57	2.86e−14/0.00	8.81/8.82	22(22)/23(23)
mivN	5	504	3383.45/3832.93	3834.69/3831.44	0.595/0.0388	1.62/1.60	0(0)/1(1)
nuoL3	5	499	5006.16/4978.97	5011.37/4977.19	1.56e−12/1.44e−15	4.58/4.49	9(8)/10(10)
perM	5	351	2619.88/2619.43	2621.64/2617.94	0.638/0.0247	1.78/1.80	0(0)/2(0)
pgpA	5	198	1541.27/1539.29	1542.65/1538.91	0.138/0.0238	2.93/2.23	1(0)/1(1)
pol	23	947	9394.05/9363.96	9405.74/9365.88	8.52e−14/0.00	1.31/1.30	6(6)/10(13)
RfaL	5	403	3964.89/3955.34	3970.38/3955.44	7.16e−05/3.23e−7	3.46/3.46	2(1)/4(3)
tax	20	181	895.50/892.02	895.50/892.02	0.0309/0.0309	0.13/0.13	181(0)/181(21)
TrbL-VirB6_2	5	657	5492.55/5492.52	5301.23/5286.43	0.976/3.74e−7	2.12/2.10	0(0)/1(0)
TrbL-VirB6_3	5	938	8305.65/8288.36	8307.06/8269.09	3.09e−8/0.00	3.06/3.02	3(2)/18(11)
vif	29	192	3393.83/3367.86	3400.45/3370.66	2.29e−06/1.16e−13	2.90/2.91	10(8)/10(10)

NOTE.—N_t: number of taxa, N_c: sequence length in number of codons, −lnL: −log likelihood for each nested model pair, p-value of the likelihood ratio test for the presence of positive selection, TTL: total tree length estimated under M2a/M8, N_s: number of sites classified to under positive selection using a posterior probability threshold of 0.95 under M2a/M8 for NEB(BEB).

Table 3. Spearman Rank Correlations between Site Posterior Probabilities for Each Method of Classification.

	Regular		*Irregular*	
	Mean	SD	Mean	SD
M2a				
NEB/BEB	0.98	0.04	0.65	0.17
NEB/SBA	0.94	0.09	0.66	0.17
BEB/SBA	0.96	0.05	0.98	0.02
M8				
NEB/BEB	0.99	0.01	0.84	0.30
NEB/SBA	0.96	0.04	0.81	0.27
BEB/SBA	0.98	0.03	0.98	0.02

NOTE.—The mean and standard deviation (SD) of the correlations are for real genes displaying *regular* and *irregular* estimation properties.

posterior of positive selection is. For the real data, these confidence intervals differed substantially between M2a and M8, highlighting differences between the two modeling frameworks. For sites having a posterior of at least 0.9 under one or more methods, the M8 confidence intervals for those sites were never wider than the corresponding M2a intervals (table 6). This result reflects broad differences between the MLE distributions obtained under these two models; MLE distributions under M8 tend to be tighter, and more likely located away from a boundary (supplementary figs. S10 and S11, Supplementary Material online). We believe this represents empirical support for the commonly held notion that M8 is more powerful than M2a (Wong et al. 2004). However, this relationship should not be assumed to hold when the MLEs are poorly estimated. Confidence interval widths were at the maximum (1.0) for both M8 and M2a in three of the five genes representing *irregular* estimation. These findings highlight the importance of (1) inspecting bootstrap distributions to gain insights into the challenges posed by the data in hand, and (2) using SBA to accommodate MLE uncertainties (especially when they are poorly estimated).

Lastly, we interpret our results for the *tax* gene of the human T-cell lymphotropic virus. This gene warrants special attention because it has a highly unusual site-pattern distribution, extreme MLEs, and has been employed as a boundary case in several studies of the NEB and BEB classifiers (Suzuki and Nei 2004; Yang et al. 2005). The dataset has 20 taxa and 181 sites, 158 (87%) of which are invariant across all 20 lineages. At each of the 23 variable sites, there is just one codon that differs from all the others with 21 of the 23 codon

Table 4. Number of Sites Identified to be under Positive Selection for the Real Data.

	M2a			M8		
Gene	NEB	BEB	SBA	NEB	BEB	SBA
CDH3	1/1	12/1	46/0	1/1	22/1	117/5
mivN	1/0	7/0	1/0	4/1	12/1	28/0
pgpA	1/0	4/0	4/0	5/1	5/1	17/0
tax	181/181	181/0	181/0	181/181	181/21	181/21
TrbL-VirB6_2	0/0	16/0	0/0	11/1	18/0	59/0
pol	12/6	19/6	94/4	22/10	33/13	83/16
lysin	33/22	32/22	42/5	37/23	37/23	41/11
nuoL3	18/9	18/8	85/18	19/10	20/10	83/20
RfaL	20/2	20/1	70/1	33/4	41/3	74/3
TrbL-VirB6_3	28/3	27/2	73/9	45/18	44/11	134/48
vif	13/10	13/8	31/6	15/10	19/10	37/10
β-globin	4/0	5/0	11/0	8/4	8/4	17/4
ccmF	7/1	11/1	112/0	15/3	79/5	114/5
ENAM	9/1	21/1	184/0	44/2	31/1	78/1
env	14/3	16/3	21/3	16/3	22/5	24/3
perM	4/0	6/0	0/0	6/2	6/0	36/3

NOTE.—The posterior probability thresholds are 0.5/0.95. The top genes represent *irregular* estimation, the middle *regular*, and the bottom genes are not categorized.

Table 5. Spearman Rank Correlations between Site Posterior Probabilities for Models M2a and M8.

	Regular		*Irregular*	
	Mean	SD	Mean	SD
NEB	0.98	0.04	0.81	0.13
BEB	0.99	0.01	1.00	0.01
SBA	1.00	0.00	0.99	0.00

NOTE.—The mean and standard deviation (SD) of the correlations are for real genes displaying *regular* and *irregular* estimation properties.

Table 6. Average SBA Posterior Probability Interval Widths for Sites with at Least One Method Having a Posterior Probability over 0.9.

Gene	M2a	M8	Difference
CDH3	0.95	0.46	0.49
mivN	1.00	1.00	0.00
pgpA	1.00	1.00	0.00
tax	0.87	0.31	0.56
TrbL-VirB6_2	1.00	1.00	0.00
pol	0.78	0.78	0.00
lysin	0.70	0.49	0.20
nuoL3	0.26	0.21	0.05
RfaL	0.68	0.48	0.19
TrbL-VirB6_3	0.66	0.10	0.57
vif	0.36	0.14	0.21
β-globin	1.00	0.00	1.00
ccmF	1.00	0.49	0.51
ENAM	0.53	0.43	0.10
env	0.51	0.27	0.24
perM	0.91	0.14	0.77

NOTE.—The top genes represent *irregular* estimation properties, the middle *regular*, and the bottom genes are not categorized.

changes coding for a different amino acid. This atypical site-pattern distribution corresponds to a relatively large number of nonsynonymous substitutions over very short branch lengths (mean branch length: 0.0064 under both M2a and M8). A very high probability of positive selection (i.e., large values for both the $p_{\omega>1}$ and $\omega_{>1}$ parameters) is required to account for the nonsynonymous substitutions when the branch lengths are so short. In fact, both models M2a and M8 estimate 100% of the sites to be in the $\omega > 1$ class. This result belies that fact that considerable instability is associated with those parameter estimates, as revealed by bootstrapping (supplementary figs. S10 and S11, Supplementary Material online). Since NEB ignores parameter value uncertainty, it must assign a conditional posterior probability of $\omega > 1$ (Pr) equal to 1.0 for all sites, including those that are invariant. In contrast, the site posteriors for BEB and SBA were similar and depended on the site patterns (supplementary table S2, Supplementary Material online). As expected, the SBA signal for positive selection was strongest at the 21 sites with nonsynonymous changes (M2a: $0.87 < Pr < 0.89$; M8: $0.99 < Pr < 0.99$), as compared with all other sites (M2a: $0.55 < Pr < 0.60$; M8: $0.76 < Pr < 0.80$). The SBA confidence intervals under M8 revealed that the estimates of *Pr* for the 21 sites with a nonsynonymous change were more reliable (average width: 0.028) than for the invariant sites (average width: 0.418). We suggest this result is appropriate for these data. Almost all the signal in this dataset is contained in those 21 sites, and it is difficult to reconcile this amount of nonsynonymous change over such short branches without strong positive selection. Moreover, when branch lengths are very short, an invariant site can only be viewed as carrying no signal about whether the ω value would be small or large over longer evolutionary periods. This leads to very wide 95% SBA *Pr* confidence intervals for these sites.

Discussion

We have presented an approach, based on an unconventional use of the nonparametric bootstrap, for evaluating MLE instabilities and improving site-specific inference of positive selection. For any given site in an alignment, conclusions about positive selection are based on the aggregation and distributions of many estimates of ω and many posterior probabilities. An important step in our approach involves smoothing the bootstrap distributions of the parameter estimates using techniques borrowed from kernel density estimation. This step is critical for overcoming instabilities in parameter estimation. Kernel smoothing also has the benefit of reducing computational costs relative to procedures that use full bootstrap sampling to obtain comparable numbers of MLEs.

Application of BEB, NEB, and SBA using models M2a and M8 to 100 simulated datasets in each of 10 different simulation scenarios showed that, under difficult simulation conditions when regularity conditions have not been met, NEB often poorly controls false positive classification of sites, even when the number of taxa is large. This is in contrast to past recommendations, which suggested NEB does well at controlling false positive rates when analyzing datasets with many taxa and long sequences (Yang et al. 2005). By accounting for variability of estimation, both BEB and SBA achieve

better control of the false positive rates. However, SBA provided consistently better control under M8 when there was mild model misspecification (studies 5–8 under in table 1), and this was unaffected by pre-screening via the LR test. We note that all real data are expected to be affected, to some degree, by model misspecification.

By accounting for variability of estimation, both BEB and SBA achieve good power relative to NEB. This is evident from the ROC curves, where the curves for BEB and SBA tend to be closer to the expected upper limit. Some of the simulation results suggest that M2a is a better-performing model than M8. For instance, M2a gave (1) ROC curves closer to the expected upper bound in some cases (fig. 3) and (2) lower false positive rates (table 1). This may, however, be a consequence of the simulations conditions being more suitable for M2a than M8. For example, in studies 3 and 7, half the sites were simulated with $\omega = 1$, and M2a has a site class with $\omega = 1$ fixed, while M8 does not. On the other hand, considering sites with larger posteriors in the real data analysis, the 95% posterior confidence intervals were usually narrower (and never wider) for M8 than M2a. This supports previous results that suggest M8 has more power to detect sites under positive selection (Wong et al. 2004). The β-globin gene serves as a good example. Of the five sites in this gene where either NEB or BEB gave a posterior of at least 0.9, the SBA confidence interval widths were all 1 for M2a, but averaged 0.129 for M8. Moreover, the $\omega_{>1}$ parameter distributions tended to be wider for M2a than M8, particularly for the genes that displayed properties suggesting regularity conditions were met. This is probably because the beta distribution used by M8 to model $\omega < 1$ has more flexibility in real data conditions compared with an M2a model with the same number of parameters.

An appealing attribute of BEB, relative to SBA, is its limited use of computational resources. Each SBA bootstrap analysis may use similar computational resources as BEB does for the one original dataset. However, SBA's greater computational requirements is a trade-off for a more rigorous assessment of the parameter estimation. For example, SBA adjusts for the uncertainty in all model parameters, including branch lengths, while BEB does not. A new BEB implementation that integrated over branch lengths would require costlier techniques because numerical integration does not scale well with higher dimension. Moreover, because SBA estimates each set of bootstrap parameters independently, they can be estimated in parallel. On a computing cluster with as many cores as bootstrap samples generated, the wall-clock times for BEB and SBA are comparable.

There are a limited number of BEB implementations for different models. In contrast, it is comparatively trivial to apply SBA to new models once the basic capacity for bootstrapping and parameter smoothing are in place. This could facilitate the application of SBA to a wider variety of inference problems in molecular evolution than has occurred with BEB. SBA for the popular branch-site codon model A (Yang and Nielsen 2002; Zhang et al. 2005) was implemented as a demonstration of the feasibility of SBA implementations for new models. A new, preliminary implementation, which was completed within a few hours, can be found at https://github.com/Jehops/codeml_sba. An overview of the analysis of the *NR1D1* gene (Baker et al. 2016) under SBA can be found in the supplementary file S2, Supplementary Material online.

There are useful by-products of the SBA approach for classifying sites. The histograms of the distributions of the MLEs over bootstrap samples provide insight into the degree of irregularity of the estimation. For several of the datasets, most notably the *tax* gene dataset, these histograms provided a clear indication that the MLEs were unstable. In such cases, site classifications should be accepted with caution. Even when regularity conditions have been met, the confidence intervals of the posteriors provide an additional tool for assessing the certainty about the strength of the signal for positive selection at an individual site. We suggest that future analyses of real data should include both visual inspection of bootstrap distributions and reporting of SBA-derived confidence intervals of the posterior probabilities associated with positive selection.

Bootstrapping has been shown to provide effective adjustments to EB methods in other settings. For example, Laird and Louis (1987) studied the application of bootstrapping with EB methods for random effects models where both the observations and random effects distributions were Gaussian. They argued that confidence intervals produced from bootstrap posteriors were frequently narrower than they should be and that bootstrap averaging helped to ameliorate problems. They speculated that bootstrapping would produce good EB inferences for a broad class of EB problems. In a prediction setting, a procedure that aggregates predictors generated from bootstrap replicates was proposed by Breiman (1996), which was shown to move some unstable predictors closer to optimality. The bagging procedure used in that paper is equivalent to using the median posterior to classify sites under SBA. Our experiments (data not shown) indicated that the average is a better measure of the middle of the distribution of site posterior probabilities.

While using the data in hand to account for errors in MLE estimation is helpful for detecting sites under positive selection, refinements of the SBA approach are warranted. Like other approaches, we have avoided the difficult process of calibrating for type I errors in real data. Choosing an optimal bandwidth parameter for smoothing a distribution is also a difficult process. Under-smoothing will leave spurious bumps and irregularities in the distribution and over smoothing will remove useful information and increase bias. There are different theoretical suggestions for the size of the bandwidth parameter, but these can be challenging to apply as they may depend on the unknown density (Venables and Ripley 2013, p. 176). SBA uses bootstrap distributions to highlight problems when MLEs fall on or close to their boundaries. We are hopeful that a penalized likelihood approach, which can push such estimates to the interior of the parameter space, will be helpful. Bootstrapping does well to accommodate the variance in a parameter estimate, however, when estimates are very small, the variance, even under bootstrapping, may be

underestimated. This may be a problem we encountered with the branch lengths of the *tax* gene. Some preliminary experiments show that perturbing the very small branch length estimates of the *tax* gene can cause large differences in the MLEs of the parameters of the ω distribution. This suggests that applying kernel smoothing to parameters other than those defining the ω distribution may be helpful.

SBA can be applied to a wide variety of problems in molecular evolution where uncertainties or instabilities in MLEs impact inference based on empirical Bayes. Examples where the method can be directly applied, with little or no modification, include: classification of sites into general rate categories (Mayrose et al. 2004), identification of positively selected sites in non-coding DNA (Haygood et al. 2007), identification codon sites subject to episodic change in selection pressure (Yang and Nielsen 2002), detection of Type-I functional divergence in protein sequences (Gaston et al. 2011), detection of amino acid sites having shifts in the pattern of exchangeabilities (Le et al. 2012), and detection of amino acid sites evolving under a covarion-like evolutionary process (Penn et al. 2008). With some modification, SBA could be applied to the task of ancestral state reconstruction. As the field moves towards increasingly more complex models, there will be increasing demand for methods such as SBA that can account for parameter-estimate uncertainties.

Theory and Methods

Markov Models of Codon Evolution

Consider an alignment of DNA sequences with n codon sites and denote the codons in the sequences at site h $(h = 1, \ldots, n)$ as x_h, the site pattern at site h. The models considered here are described in (Nielsen and Yang 1998) and define the relative, instantaneous substitution rate between codon i and j $(i \neq j)$ at site h as

$$Q_{ij} \propto \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition} \end{cases} \quad (1)$$

where Q is the rate matrix of a continuous-time, stationary, time-reversible Markov process. The π_j parameters above are the stationary frequencies of codon j and the κ parameter is the transition to transversion rate ratio. The ω parameter, which has an interpretation as the nonsynonymous to synonymous rate ratio, is the key parameter for the inference of positively selected sites (Yang 2006, pp. 48–68). The transition probability matrix, $P(t)$, which gives the probabilities of state changes over time, t, relates to the rate matrix, Q, by $P(t) = e^{Qt}$. For a phylogenetic tree with branch lengths, the likelihood of the data at codon site h given the parameters θ, $f(x_h|\theta)$, can be calculated using (1) and Felsenstein's pruning algorithm (Felsenstein 1973). Because sites are assumed to evolve independently, the log likelihood for a sequence alignment with more than one site pattern ($n > 1$) is the sum of the site log likelihoods, $\ell = \log(L) = \sum_{h=1}^{n} \log\{f(x_h|\theta)\}$.

To account for variability in selection pressure across sites, ω is usually allowed to vary. The models we consider are a subset of the models described in Yang et al. (2000a, 2000b), which assume the value of ω at site h comes from some distribution. To avoid difficulties applying the pruning algorithm, this distribution is always discrete with weights $p_1, \ldots, p_k$ on $\omega_1 \ldots, \omega_k$ values. With k classes, each with an estimate of the ω ratio and corresponding weight, the likelihood the data at site h then becomes $f(x_h|\theta) = \sum_{i=1}^{k} p_i f(x_h|\omega_i, \psi)$, where ψ denotes the model parameters other than those describing the ω distribution.

Bayes formula is used to calculate a posterior probability that a given site evolved under site class i with $Pr(\omega^{(h)} = \omega_i|x_h, \psi) = p_i f(x_h|\omega_i, \psi) / \sum_{j=1}^{k} p_j f(x_h|\omega_j, \psi)$. The NEB approach fails to account for sampling errors in any of the parameters estimated by ML. To accommodate the uncertainties in the parameters of the ω distribution, Yang et al. (2005) used a hierarchical BEB approach by assigning prior probabilities to these parameters.

Bootstrap Methods to Adjust for Uncertainty

To construct confidence intervals for a parameter, θ, and correct bias, Efron (1979) devised the bootstrap. A bootstrap sample, $\boldsymbol{x}^*$, is obtained by drawing the values, $x_1^*, \ldots, x_n^*$, with replacement from a random sample, $\boldsymbol{x}$. For each of $b = 1, \ldots, B$ bootstrap samples we calculate the bootstrap estimate, $\hat{\theta}^{*b}$. The distribution that assigns mass 1/B to each of the $\hat{\theta}^{*b}$ is the bootstrap distribution of $\hat{\theta}$. Bootstrap distributions are commonly used with phylogenetic data to test the topology of a proposed tree. We applied the bootstrap to site patterns in a sequence alignment to adjust for the uncertainty in parameter estimates in EB classification. The procedure is illustrated in figure 1:

(1) From an alignment of protein coding DNA sequences, $\boldsymbol{x}$, with n codon sites, randomly sample site patterns with replacement to obtain a bootstrap sample, $\boldsymbol{x}^{*b}$, with n sites.
(2) Estimate the MLEs, $\hat{\theta}^{*b}$, for bootstrap sample $\boldsymbol{x}^{*b}$.
(3) Use $\hat{\theta}^{*b}$ and $\boldsymbol{x}$ to calculate posterior probabilities, $Pr_h(\omega > 1|x_h, \hat{\theta}^{*b})$, that each site, h, is under positive selection.
(4) Repeat steps 1 through 3 B times to calculate B sets of posterior probabilities for each codon site.
(5) Calculate an aggregate posterior probability that each site is under positive selection by, e.g., averaging posterior probabilities over bootstrap replicates, $\sum_{b=1}^{B} Pr_h(\omega > 1|x_h, \hat{\theta}^{*b})/B$.

A preliminary implementation of the SBA method supporting codon models M2a, M8, and branch-site model A, built upon the codeml application from the PAML package (Yang 2007), can be found at

https://github.com/Jehops/codeml_sba.

Kernel Smoothing to Approximate the Bootstrap Distribution

Kernel smoothing (Akaike 1954; Rosenblatt et al. 1956; Parzen 1962; Wand and Jones 1994) is class of nonparametric techniques that can improve estimation of a distribution. The kernel density estimator for a continuous density f,

$$\hat{f}(x;h) = (nh)^{-1}\sum_{i=1}^{n} K([x - X_i]/h),$$

includes a kernel density (probability) function, K, to locally average or smooth observations and the amount of smoothing is controlled by a bandwidth parameter, h. For small h, each of the $h^{-1}K([x - X_i]/h)$ contributions are large only for x close to some X_i giving rise to a bumpy distribution, whereas for h large the $h^{-1}K([x - X_i]/h)$ contributions overlap giving a much smoother distribution (Silverman and Young, 1987). We used kernel density estimation to create smoothed bootstrap distributions for the p parameters of the ω distributions under models M2a and M8 using a uniform kernel.

Kernel density estimation requires a bandwidth parameter as input. One method for determining h is using leave-one-out cross validation,

$$\hat{f}_{(-k)}(x;h) = (n-1)^{-1}h^{-1}\sum_{i\neq k} K([x - X_i]/h)$$

(Venables and Ripley 2013, p. 184). In this approach, h is chosen to maximize the sum of the logged density estimates $\sum_k \log \hat{f}_{(-k)}(x_k;h)$, where $\hat{f}_{(-k)}(x;h)$ is the kernel density estimate constructed from all of the x_i except x_k. However, our experiments using leave-one-out likelihood to choose an optimal bandwidth parameter for the p parameters of M2a and M8 merely resulted in smoothed estimates of the biased bootstrap distributions. To obtain conservative estimates of the p parameters that suppressed the influence of instabilities we chose to over smooth by using a bandwidth parameter of $h = 0.4$ for all applications of SBA.

Adding kernel smoothing to the bootstrap algorithm increases the number of parameter estimates used in step 5 of the unsmoothed algorithm by sampling from a smoothed bootstrap distribution. The adjustment is in step 2 of the algorithm. The ML parameters estimated from bootstrap sample b, $\hat{\theta}^{*b}$, are replaced by θ^{sb} sampled from the smoothed bootstrap distribution. The rest of the algorithm proceeds as in the unsmoothed version, but using θ^{sb} in place of $\hat{\theta}^{*b}$.

For model M8, the step 2 adjustment is as follows. For each $\hat{\theta}^{*b}$, $p^{sb}_{\omega<1}$ samples are repeatedly drawn from a univariate uniform distribution centered at $\hat{p}^{*b}_{\omega<1}$ with width $2h$. If necessary, the minimum and maximum points of the distribution are truncated to 0 and 1. Let θ^{sb} denote $\hat{\theta}^{*b}$ with $p^{sb}_{\omega<1}$ replacing $\hat{p}^{*b}_{\omega<1}$ ($p^{sb}_{\omega>1} = 1 - p^{sb}_{\omega<1}$). The same procedure is used under model M2a, however, with three weight parameters, the sampling is done on a bivariate uniform distribution with the following additional restrictions:

i) $p^{sb}_{\omega<1} + p^{sb}_{\omega=1} \leq 1$,
ii) $(\hat{p}^{*b}_{\omega<1} - h) \leq p^{sb}_{\omega<1} \leq (\hat{p}^{*b}_{\omega<1} + h)$, and
iii) $(\hat{p}^{*b}_{\omega=1} - h) \leq p^{sb}_{\omega=1} \leq (\hat{p}^{*b}_{\omega=1} + h)$.

As with M8, if necessary, the minimum and maximum points of the distribution are truncated at 0 and 1, and $p^{sb}_{\omega>1} = 1 - p^{sb}_{\omega<1} - p^{sb}_{\omega=1}$.

Simulation Studies

Datasets were simulated using *EvolverNSSites* from the PAML 4.8a package (Yang 2007) and *Indelible* (Fletcher and Yang 2009) following some of the settings described in Wong et al. (2004). To compare the relative performance of BEB, NEB, and SBA for predicting sites under positive selection, ten different simulation studies, divided into three scenarios, were used. Table 1 gives an overview of the ω distributions used to simulate the data. The *Correct Model Scenario* included four simulation studies where the nuisance parameters, $\kappa = 1$ and $\pi_i = 1/61$, matched the fitted model. The *Mild Misspecification* and *Heavy Misspecification* scenarios included four simulation studies with mild misspecification and two studies with heavy misspecification of the fitted model, respectively. The data in the *Mild Misspecification Scenario* was simulated using $\kappa = 8$ and empirical codon frequencies derived from application of the general time-reversible model (Yang 2006, p. 33) to the *TrbL-VirB6-3* plasmid conjugative transfer protein of *Rickettsia*. In the fitted model, κ was estimated, while the misspecification was introduced by using F3×4 (expected codon frequencies calculated using the nucleotide frequencies at the three codon positions). For the *Heavy Misspecification Scenario*, study 9 used the heavily biased codon frequencies from the *Drosophila GstD1* gene and $\kappa = 8$ to simulate the data. In study 10, there were two heterogeneous classes of sites. Half the sites were simulated using equal codon frequencies, $\kappa = 1$, and $\omega = 0.5$, while the other half with the *Drosophila GstD1* gene codon frequencies, $\kappa = 8$, and $\omega = 1$. For both studies in this scenario, analysis was carried out using a single set of codon frequencies (set equal to 1/61) and a single κ parameter estimated for all sites in the data set. For all studies in the three scenarios, 100 alignments, each having 500 codons, were simulated with the same 5-taxon tree from Wong et al. (2004). The studies in the *Correct Model Scenario* were repeated under model M2a with the 30-taxon tree from the same paper.

Real Data Analysis

Table 2 describes the real data sequences we analyzed under models M2a and M8 using NEB, BEB, and SBA. Of the 16 genes, eight code for transmembrane proteins in *Rickettsia* (*ccmF*, *mivN*, *perM*, *pgpA*, *RfaL*, *TrbL-VirB6_2*, and *TrbL-VirB6_3*) and were previously analyzed in Bao et al. (2008). Three genes from the HIV-1 virus (*env pol*, and *vif*) and a β-globin gene were described and analyzed in Yang et al. (2000a), two primate genes (*CDH3* encoding cadherin and *ENAM* encoding enamelin), a lysin gene from Yang et al. (2000b), and the *tax* gene from the human T-cell lymphotrophic virus (HTLV) that was analyzed by Suzuki and Nei (2004). All data is available from https://github.com/Jehops/sba_real_data.

Supplementary Material

Supplementary figures S1–S11, tables S1–S2, and files S1–S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

References

Akaike H. 1954. An approximation to the density function. *Ann Inst Stat Math*. 6:127–132.

Anisimova M, Bielawski J, Dunn K, Yang Z. 2007. Phylogenomic analysis of natural selection pressure in Streptococcus genomes. *BMC Evol Biol*. 7:154.

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.

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

Aris-Brosou S. 2003. How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics* 19:618–624.

Baker JL, Dunn K, Mingrone J, Wood BA, Karpinski BA, Sherwood CC, Wildman DE, Maynard TM, Bielawski JP. 2016. Functional divergence of the nuclear receptor nr2c1 as a modulator of pluripotentiality during hominid evolution. *Genetics* 203:905–922.

Bao L, Gu H, Dunn KA, Bielawski JP. 2008. Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. *Mol Biol Evol*. 25:1995–2007.

Bickel PJ, Doksum KA. 2006. Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I. 2nd ed. Boca Raton: CRC Press.

Bielawski JP, Yang Z. 2005. Maximum likelihood methods for detecting adaptive protein evolution. In: Nielsen R, editor. Statistical Methods in Molecular Evolution. New York: Springer, pp. 103–124.

Breiman L. 1996. Bagging predictors. *Mach Learn* 24:123–140.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*. 16:1457–1465.

Davison AC, Hinkley DV. 1997. Bootstrap Methods and their Application, Vol. 1. Cambridge: Cambridge University Press.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Ann Stat* 1–26.

Efron B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans, Vol. 38. Philadelphia: SIAM.

Efron B, Tibshirani RJ. 1994. An Introduction to the Bootstrap. Boca Raton: CRC Press.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*. 25:471.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A*. 94:7712–7718.

Fletcher W, Yang Z. 2009. Indelible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 26:1879–1888.

Gaston D, Susko E, Roger AJ. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics* 27:2655–2663.

Ge G, Cowen L, Feng X, Widmer G. 2008. Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Comp Funct Genom*. 2008:879023.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39:1140–1144.

Huelsenbeck JP, Dyer KA. 2004. Bayesian estimation of positively selected sites. *J Mol Evol*. 58:661–672.

Kalbfleisch J. 1985. Probability and Statistical Inference: Volume 2: Statistical Inference. Springer Texts in Statistics. New York: Springer.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624.

Laird NM, Louis TA. 1987. Empirical Bayes confidence intervals based on bootstrap samples. *J Am Stat Assoc*. 82:739–750.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. 29:2921–2936.

Lemey P, Minin VN, Bielejec F, Pond SLK, Suchard MA. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics* 28:3248–3256.

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

Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*. 21:1781–1791.

Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol*. 51:729–739.

Nielsen R, Huelsenbeck JP. 2002. Detecting positively selected amino acid sites using posterior predictive p-values. *Pac Symp Biocomput*. 7:576–588.

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.

Parzen E. 1962. On estimation of a probability density function and mode. *Ann Math Stat*. 33:1065–1076.

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol*. 4:e1000214.

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

Rosenblatt M, et al. 1956. Remarks on some nonparametric estimates of a density function. *Ann Math. Stat* 27:832–837.

Scheffler K, Murrell B, Pond SLK. 2014. On the validity of evolutionary models with site-specific parameters. *PloS One* 9:e94534.

Silverman B, Young G. 1987. The bootstrap: to smooth or not to smooth? *Biometrika* 74:469–479.

Suzuki Y. 2004. New methods for detecting positive selection at single amino acid sites. *J Mol Evol*. 59:11–19.

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

Suzuki Y, Nei M. 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.

Venables WN, Ripley BD. 2013. Modern Applied Statistics with S-PLUS. New York: Springer.

Wand P, Jones C. 1994. Kernel Smoothing. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. New York: Taylor & Francis.

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. 2005. The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci U S A*. 102:3179–3180.

Yang Z. 2006. Computational Molecular Evolution. Oxford (United Kingdom): Oxford University Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*. 46:409–418.

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

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

Yang Z, Swanson WJ, Vacquier VD. 2000b. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol*. 17:1446–1455.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

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