Identifying sites under positive selection with uncertain parameter estimates

Stéphane Aris-Brosou

Abstract: Codon-based substitution models are routinely used to measure selective pressures acting on protein-coding genes. To this effect, the nonsynonymous to synonymous rate ratio (dN/dS = ω) is estimated. The proportion of amino-acid sites potentially under positive selection, as indicated by ω > 1, is inferred by fitting a probability distribution where some sites are permitted to have ω > 1. These sites are then inferred by means of an empirical Bayes or by a Bayes empirical Bayes approach that, respectively, ignores or accounts for sampling errors in maximum-likelihood estimates of the distribution used to infer the proportion of sites with ω > 1. Here, we extend a previous full-Bayes approach to include models with high power and low false-positive rates when inferring sites under positive selection. We propose some heuristics to alleviate the computational burden, and show that (*i*) full Bayes can be superior to empirical Bayes when analyzing a small data set or small simulated data, (*ii*) full Bayes has only a small advantage over Bayes empirical Bayes with our small test data, and (*iii*) Bayesian methods appear relatively insensitive to mild misspecifications of the random process generating adaptive evolution in our simulations, but in practice can prove extremely sensitive to model specification. We suggest that the codon model used to detect amino acids under selection should be carefully selected, for instance using Akaike information criterion (AIC).

Key words: codon substitution models, empirical Bayes, Bayes empirical Bayes, full Bayes, ROC curves, AIC.

Résumé : Les modèles de substitutions de codons sont couramment utilisés pour mesurer les pressions de sélection qui agissent sur les gènes codant pour des protéines. Pour ce faire, le rapport des taux de substitutions non synonymes à celui des substitutions synonymes (dN / dS = ω) est calculé. La proportion d'acides aminés susceptibles d'être sous sélection positive, indiquée par ω > 1, est inférée par l'ajustement d'une distribution qui permet certain sites d'avoir ω > 1. Une approche empirique Bayes ou Bayes empirique Bayes est ensuite utilisée pour identifier ces sites, soit, respectivement, en négligeant les erreurs d'échantillonnage des estimateurs de maximum de vraisemblance, soit en en tenant compte. Ici nous bâtissons sur une approche hiérarchique Bayes récemment développée afin d'inclure des modèles de codons à puissance élevée et un taux de faux positifs raisonnable. Nous proposons des heuristiques pour accélérer les calculs, et montrons que (*i*) l'approche hiérarchique Bayes est supérieure à une approche empirique Bayes sur l'analyse d'un petit jeu de données ou de données simulées mais (*ii*) n'a qu'un faible avantage par rapport a une approche Bayes empirique Bayes et (*iii*) les méthodes Bayesiennes semblent robustes aux misspécifications légères du processus stochastique générant l'évolution adaptative dans nos simulations, mais en pratique sont extrêmement sensibles à la spécification du modèle. Nous suggérons que le modèle de codon utilisé pour détecter les acides aminés sous sélection soit soigneusement sélectionné, par exemple en utilisant AIC.

Mots clés : modèles de substitutions de codons, Bayes empirique, Bayes empirique Bayes, modèles hiérarchiques, courbes ROC, AIC.

Introduction

The identification of amino-acid sites potentially under positive selection has attracted some recent attention (e.g., Suzuki and Nei 2004; Wong et al. 2004; Yang et al. 2005). The criterion used to detect positive selection in protein-coding genes is based on a comparison of nonsynonymous (dN) and synonymous (dS) rates. When the nonsynonymous rate is greater than the synonymous rate, the rate ratio dN/dS = ω is >1, which is interpreted as evidence of the action of positive selection (e.g., Yang 2001). Conceptually, the simplest approach to detecting sites under positive selection is a site-by-site or sitewise approach, either counting numbers of substitutions on a phylogeny (e.g., Suzuki and Nei 2004) or using a maximum-likelihood estimation (Kosakovsky Pond and Frost 2005*b*; Massingham and Goldman 2005). Whereas sitewise approaches are quite powerful on data sets with a large number of sites (Kosakovsky Pond and Frost 2005*b*; Massingham and Goldman 2005), they generally disregard sampling errors related to parameter estimates. This is a potential issue for small data sets, since maximum-likelihood estimates (MLEs) can have large sampling errors. Unlike sitewise models, the most common approach relies on random-effect models that consider the entire sequence align-

Received 21 December 2005. Accepted 5 March 2006. Published on the NRC Research Press Web site at http://genome.nrc.ca on 11 August 2006.

Corresponding Editor: B. Golding.

S. Aris-Brosou. Department of Biology, University of Ottawa, 30 Marie Currie, Ottawa, ON K1N 6N5, Canada (e-mail: sarisbro@uottawa.ca).

 doi:10.1139/G06-038

ment, or partitions thereof, and lend themselves easily to accommodating uncertainties.

In this latter approach, a codon model specifies the evolution of protein-coding sequences. The model, originally described by Goldman and Yang (1994), denoted M0, assumes that each site of the protein evolves under the same ω. Nielsen and Yang (1998) and Yang et al. (2000) extended this basic model to incorporate a random-effect model that allows ω to vary among sites, thereby making it possible to detect sites potentially under positive selection. This is achieved by superimposing on the codon model a probability distribution that describes how ω varies among sites. A number of such distributions or random models have been proposed (Yang et al. 2000; Swanson et al. 2003; Wong et al. 2004). Evidence of positive selection is then detected by comparing a null model that does not allow for sites under positive selection with a more general model that allows for such sites. The comparison is performed by means of a likelihood ratio test (LRT). Two LRTs appear to have satisfying power and reasonable false-positive rates (Wong et al. 2004). The first compares the null model (M1a), which assumes 2 discrete categories of sites in proportion p_0 and $p_1 = 1 - p_0$, where $0 < \omega_0 < 1$ and $\omega_1 = 1$; the alternative model (M2a) adds a third category of sites, $\omega_2 > 1$, estimated as a free parameter. The second useful test compares model M8, which assumes a beta distribution for ω in the interval (0,1) plus a point mass at $\omega_2 = 1$, with model M8a, where ω_2 is estimated as a free parameter. If the LRT is significant and if the estimated ω_2 is >1, evidence of positive selection is found, although we do not know which sites caused the null model to be rejected, that is, which sites are potentially under positive selection.

In this framework, identification of such sites is done a posteriori (after the likelihood analysis and after the LRT), by means of an empirical Bayes procedure (Nielsen and Yang 1998; Yang et al. 2000). This consists of computing for each site i with data-column x_i in an alignment, its posterior probability of belonging to rate category k:

$$[1] \qquad p(\omega^{(i)} = \hat{\omega}_k | x_i) = p(x_i | \hat{\omega}_k, \hat{\lambda})\hat{p}_k / \Sigma_j p(x_i | \hat{\omega}_j, \hat{\lambda})\hat{p}_j$$

The vector λ contains the nuisance parameters specific to the substitution model, such as the branch lengths and the transition to transversion-rate ratio, and p_k represents the prior assumption on the parameters of the random model. In this approach, the prior assumption, estimated from the data, reduces to the frequency of rate category k when a discrete model, such as M2a, is used. The *empirical* part of the procedure comes from plugging MLEs (denoted by hats) into the Bayes inversion formula (eq. 1). If the posterior probability that site i belongs to a rate category k with $\hat{\omega}_k > 1$ is greater than a cut-off threshold, e.g., 0.95, then site i is inferred to be under positive selection.

An important shortcoming of the empirical Bayes approach is that in none of the model parameters — λ, p_k, or ω — is uncertainty accounted for (Yang et al. 2005). As noted above for sitewise likelihood models, this is a concern, especially in the case of small data sets where MLEs can have large sampling errors. This makes the traditional empirical Bayes approach potentially unreliable (Anisimova et al. 2002). To date, 2 remedies have been proposed. The first solution is to adopt a hierarchical or full-Bayes (FB) approach, where all the model parameters on the real line and the tree topology are assumed to follow some prior distributions (Huelsenbeck and Dyer 2004). Sites under positive selection are then identified as those for which the posterior probability of belonging to a rate category k with $\omega_k > 1$ is the highest. Uncertainty over all model parameters is integrated over uninformative or vague prior distributions. However, the statistical properties of the FB approach have never been evaluated, and the method has not been implemented in the most useful models, like M2a and M8a, which have satisfying power and reasonable false-positive rates (Yang et al. 2005). It is also unclear how high this posterior probability should be to be able to make a fair comparison between the empirical and the FB approaches. Posterior distributions can be summarized in a number of ways, using their means or their modes. Huelsenbeck and Dyer (2004) empirically found that the sites identified using the median were more similar to those derived from the empirical-Bayes approach, and based their inference on the median rather than the mean. However, they found a puzzling lack of correspondence between this summary of FB posterior probabilities and their empirical-Bayes counterparts.

An alternative solution, one frequentist in essence but that tries to mimic a FB approach (Deely and Lindley 1981), is to use the information at the other sites in an alignment to help define the prior probability $p(\omega_k)$. This approach, called Bayes empirical Bayes (BEB), first assumes that ω_k is distributed according to a prior distribution, $p(\omega_k \mid \theta)$. In this hierarchical design, $\theta = \{\lambda, p_k, \omega\}$ is a hyperparameter vector that governs the distribution of ω_k and is distributed according to a hyperprior distribution, $p(\theta)$. Following these assumptions, the BEB posterior probability can be derived (Yang et al. 2005). What is left then is the specification of the prior and hyperprior distributions, $p(\omega_k|\theta)$ and $p(\theta)$. These can be taken as uniform (e.g., for ω_k) or "flat" Dirichlet distributions (e.g., for rate-category frequencies). One major difference with the FB approach is computational. The approximation of BEB posterior probabilities does not resort to Markov chain Monte Carlo (MCMC) samplers. Yang et al. (2005) used discrete integration. Other approximation techniques are possible, like the bootstrap procedure proposed by Laird and Louis (1987). The second major difference with the FB approach is that Yang et al. (2005) did not integrate over the entire vector of nuisance parameters θ; only the parameters of the random model for sites (p_k, ω) were integrated over. Parameters λ of the rate matrix and branch lengths were still set to their MLEs. It is unclear how ignoring these uncertainties can affect our procedures to identify sites under positive selection.

Kosakovsky Pond and Frost (2005*b*) recently performed an extensive simulation study, comparing sitewise approaches (counting and likelihood methods) with empirical-Bayes methods for detecting amino-acid sites under positive selection, and showed that both approaches had roughly similar type I and type II error rates. Yang et al. (2005) compared empirical Bayes and BEB methods, and showed the superiority of BEB when analyzing small data sets. Here, we extend those comparisons to evaluate the statistical performance of the FB approach. These are evaluated for the most useful models, M2a and M8a, not implemented by Huelsenbeck and Dyer (2004). We apply these models and

identification criteria to a small HIV-1 data set, propose and test the effect of simple computational heuristics, assess the statistical performance of frequentist and Bayesian methods using receiver operating characteristic (ROC) analysis, and finally discuss their merits and shortcomings with a special emphasis on their robustness to model specification.

Theory

Full Bayes on posterior probabilities

The goal is to assign probabilistically each site of an alignment to a rate category based on its posterior probability. Uncertainties in the parameter estimates are accounted for by performing the following integration:

$$[2] \qquad p(\omega^{(i)} = \omega_k | X) = E_{\theta|X}[p(\omega^{(i)} = \omega_k | X, \theta)]$$

which means that parameters $\theta = \{\lambda, p_k, \omega\}$ are drawn from their joint posterior distribution, approximated by constructing an MCMC sampler. Given θ, the conditional posterior probability that each site i belongs to category k is then computed according to eq. 2 (see also Huelsenbeck and Dyer 2004). We will refer to the quantity on the left-hand side of eq. 2 as the posterior mean *integrated posterior probability* (meanIPP).

The probabilistic assignment of individual sites to a rate category can be based on several criteria. The first criterion we describe follows that described by Huelsenbeck and Dyer (2004). A given site i will be likely to be under positive selection if 2 conditions are met: the meanIPP is large enough, and the posterior mean rate, $E_{\theta|X}[\omega_k]$, of this category with the largest meanIPP is greater than 1. To help comparisons with frequentist procedures, a threshold α can be chosen by calibrating the criterion so that the estimated probability of identifying sites under positive selection equals the true probability (Andrews 1970). Taking an empirical approach, Huelsenbeck and Dyer (2004) found that using the median rather than the mean resulted in sets of sites more similar to those obtained by empirical Bayes inference. We then define a measure based on largest posterior median of integrated posterior probabilities (medianIPP).

In practice, however, integrated posterior probabilities $p(\omega^{(i)} = \omega_k|X)$ can have a large variance (Huelsenbeck and Dyer 2004), and the criteria defined above might become misleading when integrated posterior probabilities are highly skewed at some sites. To help alleviate this issue, we define a 3rd criterion, based on the percentile rank of integrated posterior probabilities at a given threshold α (PR_α). The percentile rank is the proportion α of the distribution $p(\omega^{(i)} = \omega_k|X)$ that is greater than or equal to a certain threshold ϕ. As above, the criterion can be calibrated on both thresholds, but to simplify the argument, the second one (ϕ) was fixed to a posterior probability of 95%.

Full Bayes on site rates

As stated above, the objective is to identify the sites of an alignment that are potentially under positive selection, i.e., those sites for which $\omega^{(i)} > 1$. The most natural approach would be to estimate sitewise posterior mean $\omega^{(i)}$ values directly, while averaging out uncertainties about estimates of model parameters as:

$$[3] \qquad E_{\theta|X}[\omega^{(i)}] \approx \frac{1}{M} \sum_{j}^{M} \sum_{category\, k} \omega_k^{\{j\}} p^{\{j\}}(\omega^{(i)} = \omega_k | X)$$

Parameters are drawn from the posterior distribution, and the quantity $p^{(j)}(\omega^{(i)} = \omega_k|X)$ of eq. 2 is calculated at each step j of the Markov chain, which is run on the state space of $\theta = \{\lambda, p_k, \omega\}$. Massingham and Goldman (2005) and Kosakovsky Pond and Frost (2005*b*) proposed a related yet different computation that is not based on random-effect models, as it is here, but on fixed-effect models. Their models allow sitewise $\omega^{(i)}$ values to be estimated rather than their posterior means, but their models disregard uncertainties about model parameter estimates.

As above, 3 criteria can be defined. Under the first one, a site i will be under positive selection if:

$$[4] \qquad E_{\theta|X}[\omega^{(i)}] > 1$$

Hereafter, this criterion will be referred to as PMeanGO (posterior mean greater than 1). We define a related criterion based on the median of the posterior distribution of posterior means of $\omega^{(i)}$: PMedGO (posterior median greater than 1). Although these 2 criteria account for both uncertainties in parameter estimates and posterior probabilities, no thresholds are involved.

Even if these criteria explicitly take parameter uncertainties into account, they might not be reasonable ways to identify positively selected sites for the same reasons that taking $\hat{\omega}^{(i)} > 1$ in a standard maximum likelihood approach is not. An approach with tighter control might be to consider site i as potentially under positive selection if α% of the posterior distribution of $\omega^{(i)}$ is >1. As described above, α will be determined by calibration to make fair comparisons with the other criteria. Hereafter, this criterion will be denoted PDGO (posterior density greater than 1).

Materials and methods

Markov chain Monte Carlo

To approximate the Bayesian quantities defined above, an MCMC sampler of target distribution $p(\theta|X)$ with $\theta = \{\lambda, p_k, \omega\}$ was constructed as follows. A Perl script was written to (*i*) initialize starting values of the parameters of the model; (*ii*) propose a new value for one of the model parameters, say ψ, from a proposal distribution $f(.|.)$, here denoted as ψ', with new parameters chosen cyclically from among θ; and (*iii*) compute the log-likelihood externally with codeml (Yang 1997) for the proposed state and accept it with probability:

$$[5] \qquad \alpha = \min\left[1, \frac{p(\psi'|X)}{p(\psi|X)} \frac{f(\psi|\psi')}{f(\psi'|\psi)}\right] = \min\left[1, \frac{p(X|\psi')}{p(X|\psi)} \frac{p(\psi')}{p(\psi)} \frac{f(\psi|\psi')}{f(\psi'|\psi)}\right]$$

States proposed outside of the state space of a given parameter are reflected back into the defined state space. If the proposed state is accepted, set the current value of the parameter ψ to ψ'. Then return to step *(ii)* and iterate until a

large number M of samples is drawn from the target distribution, at stationarity of the chain.

Some advantages of using a Perl script and codeml externally are that likelihood computations are guaranteed to be correct (given that those in codeml are), they can be computed over the large range of state-of-the-art codon models implemented in codeml, and the entire MCMC sampler can be implemented very quickly. This was done for site models M0, M2a, M3, M7, and M8a (Yang et al. 2000; Wong et al. 2004). These models can be compared by means of the Bayes factor (e.g., Aris-Brosou 2003; Scheffler and Seoighe 2005), although we do not examine model selection in a Bayesian framework here. The limitation of using a Perl script to implement the Markov chain is, given the external calls, the mediocre computing speed.

Uncertainty about model parameters is integrated over the following prior distributions: branch lengths follow a mean–0.1 exponential distribution, the transition to transversion-rate ratio follows a uniform distribution on (0,100), ω rate ratios follow uniform distributions on (0,100), parameters p and q of the beta distribution used in M8 and M8a follow a uniform distribution on (0,15), and rate-category frequencies in models M2a, M3, and M8a follow flat Dirichlet distributions. Uniform proposal distributions $f(.|.)$ were used for all parameters, except for frequencies drawn from Dirichlet distributions. Equilibrium codon frequencies, difficult to estimate from small data sets, were calculated from the observed nucleotide frequencies (F3×4, e.g., Yang 2001). Tree topologies were set to their MLEs (HIV-1 data set) or to the generating model (simulations). A more realistic scenario for simulations might use the estimated tree topology. This should not overly affect the results presented here, because topologies have previously been shown to have little impact on procedures that identify sites under positive selection (Yang et al. 2000; Kosakovsky Pond and Frost 2005*b*). Each chain was run for 10^5 (simulations) or 10^6 (HIV-1 data set) steps, and sampled every 100 steps to reduce sample autocorrelation (thinning). Three independent chains were run for each model to help monitor convergence.

To reduce the computational burden of the MCMC sampler, we explored the possibility of starting the chain from the MLEs of the model parameters and fixing branch lengths to their MLEs. We assessed the effect of ignoring branch-length-estimation uncertainty on the criteria used to identify sites potentially under positive selection. The script implementing this sampler is available at aix1.uottawa.ca/~sarisbro (in the Downloads tab).

Data example

We reanalyzed a data set consisting of the HIV-1 envelope glycoprotein (*env*) gene V3 region from 13 HIV-1 isolates from Sweden (Leitner et al. 1997). This data set was chosen for its small size (91 codons), and because it has been used frequently as a test data set (Yang et al. 2000, 2005; Kosakovsky Pond and Frost 2005*b*; Scheffler and Seoighe 2005). Equilibrium codon frequencies were calculated using observed nucleotide frequencies (F3×4 scheme), and the tree topology was set to that used previously (Yang et al. 2000, 2005).

Simulation study

The statistical properties of the proposed criteria were assessed using simulations. To make the simulation results directly relevant to the reanalyzed HIV-1 data, 13-sequence trees, containing 91 codons, were generated using evolver (Yang 1997). The topology was the same as that used for the analyses. A total of 4 conditions were simulated, where parameters were set to the HIV-1 MLEs under the corresponding model:

1. $\omega = 0.90$ for 100% of the sites (MLEs under M0).
2. $\omega = 0.06$ for 38% of the sites, $\omega = 1$ for 44% of the sites, and $\omega = 3.63$ for 18% of the sites (MLEs under M2a).
3. $\omega = 0.00$ for 28% of the sites, $\omega = 0.73$ for 48% of the sites, and $\omega = 3.26$ for 24% of the sites (MLEs under M3, 3 discrete rate categories).
4. $\omega = 0.00$, 0.05, 0.57, 0.97, and 1.00 each for 16% of the sites, and $\omega = 3.43$ for 20% of the sites, (MLEs under M8a; the beta distribution was approximated using 5 discrete categories, rather than 10, for computation speed of simulations).

Under each condition, 100 replicates were generated using empirical codon frequencies. Each simulated condition was analyzed under models M2a and M8a, assuming an F3×4 scheme for codon frequencies. The Markov chains were started from the MLEs corresponding to the model assumed for the analysis, and run for a total of 10^5 steps, with thinning of 100. Convergence was checked by performing regressions on time-series plots of the log-likelihood values and of the $\omega^{(i)}$ parameters sampled from the posterior. This was performed on each chain independently. Posterior estimates were also checked against their respective MLEs. BEB estimates were computed with codeml (September 2004 release).

Results and discussion

Computational heuristics

We investigated 2 simple heuristics to alleviate the computational burden of an FB identification of sites under positive selection. In the first, the Markov chains are started not from a random point in the parameter space, but from the MLEs of the model parameters. The motivation here is to save on the burn-in period, which is the time required by a Markov chain to forget its initial state and reach stationarity (that is, convergence). This heuristic method is available because each parameter of the Bayes model corresponds to a parameter in the likelihood model, and all model parameters are identifiable. Conversely, this would not work with Bayes models estimating divergence times and rates, because times and rates are not identifiable in a likelihood framework, and their MLEs do not exist when the clock cannot be assumed. Convergence, always an issue with MCMC methods, is usually assessed by starting independent chains from different points in the space of model parameters and checking that each chain converges on the same distribution — the target or posterior distribution (e.g., Aris-Brosou and Yang 2002). One potential shortcoming when independent chains are started from the same point is the risk of entrapment in the same local optimum. In this case, lack of convergence on the target distribution might be difficult to diagnose. The heuristic process used here assumes that the target distribution is centered on the MLEs of the model parameters, and that the

likelihood optimization procedures do converge on the absolute optimum. In our implementation, each MCMC run is preceded by a likelihood optimization procedure (see Materials and methods). Convergence was then assessed by running the likelihood + MCMC implementation several times, each time starting from independent random points, and checking that all chains converge on the same distribution. This procedure warrants caution but appeared to perform sufficiently well in the cases we analyzed.

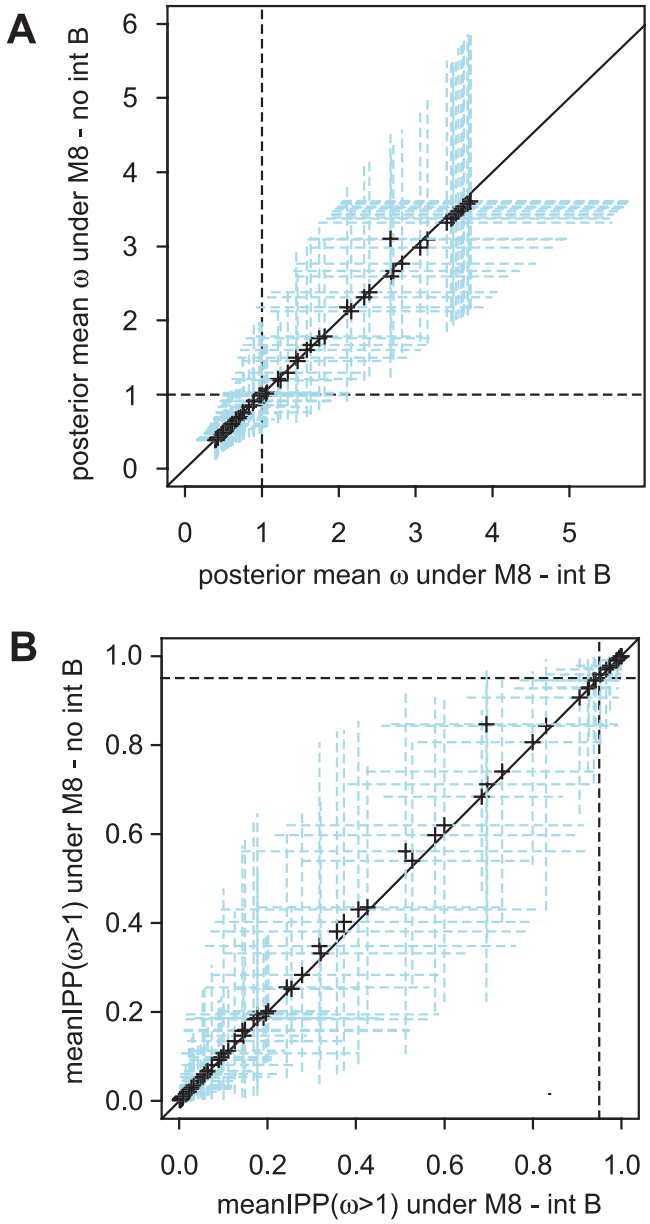
Another heuristic process tested here involves reducing the dimension of the model by fixing some model parameters to their MLEs. For the small HIV-1 data set analyzed, setting the branch lengths to their MLEs had only a negligible effect. The sitewise posterior means of ω were almost unaffected by the approximation (Fig. 1A), and estimates of meanIPPs were almost identical under the 2 integration models (Fig. 1B). In both panels of Fig. 1, the outlier is site 84K. The reason for this pattern is unclear; convergence did not seem to be the cause. Importantly, the sites identified with and without integrating over branch lengths were identical.

Saving on the burn-in period and on the integration over branch lengths has 2 distinct advantages. First, it improves mixing; fewer parameters are integrated over. Second, it reduces the running time of the MCMC samplers implemented on these parameter-rich codon models without affecting our criteria for identifying sites under positive selection. Both heuristic processes were used during our simulations: MCMC samplers were started from the MLEs of the model parameters, and branch lengths were subsequently set to the MLEs of each replicate.

HIV-1 *env* gene analysis

Quite often a model contains 2 broad classes of parameters: those in which we are directly interested, such as rates of evolution; and those that are there to make the model more realistic. Unfortunately, these nuisance parameters potentially affect the determination of the parameters of interest. When all these parameters are continuous, ignoring uncertainty about them during naïve empirical Bayes (EB) inference leads to confidence intervals about the parameters of interest that are too small (Deely and Lindley 1981). The context here is slightly different, in that we are not interested in a continuous parameter but in a binary classifier (under vs. not-under positive selection) subject to continuous nuisance parameters. As a result, it is not certain whether an analogy with the continuous case can be drawn. Should the credible set of positive sites be expected to be too small when uncertainty about model parameters is ignored? Or should it be the set of nonpositive sites that is too small? In the case of small data sets, where uncertainties are exacerbated, the BEB approach is expected to improve our inference by taking into account uncertainty about some model parameters. In the case of the well-studied HIV-1 *env* gene data set, LRTs suggested the existence of some sites under positive selection (Yang et al. 2005). At a shared posterior probability cut-off of 95%, Yang et al. (2005) found that, under M2a, both EB and BEB approaches identified the same 3 sites (28T, 66E, and 87V) (Table 1). Under the more parameter-rich model (M8a), EB identified the same 3 sites as did M2a, but BEB identified 2 extra sites (26N and 51I) (Table 1). BEB might identify larger sets of sites than EB because integrating over uncertainties about some model parameters increases the power of site-identification procedures (Yang et al. 2005). Because the only parameters integrated out are those relative to the probability distribution describing among-site variation of ω, integrating over uncertainties about all model parameters might lead to a more powerful procedure.

Fig. 1. Effect of integrating over branch lengths (int B) or not integrating over branch lengths (no int B) on sitewise posterior inference for the HIV-1 data set under model M8a: (A) effect on integrated posterior means of ω for each of the 91 codons in the alignment; (B) effect on posterior means of integrated posterior probabilities (meanIPP) for each of the 91 codons in the alignment. No int B is plotted against int B. Solid line indicates major diagonal; broken lines indicate 95% credible intervals.



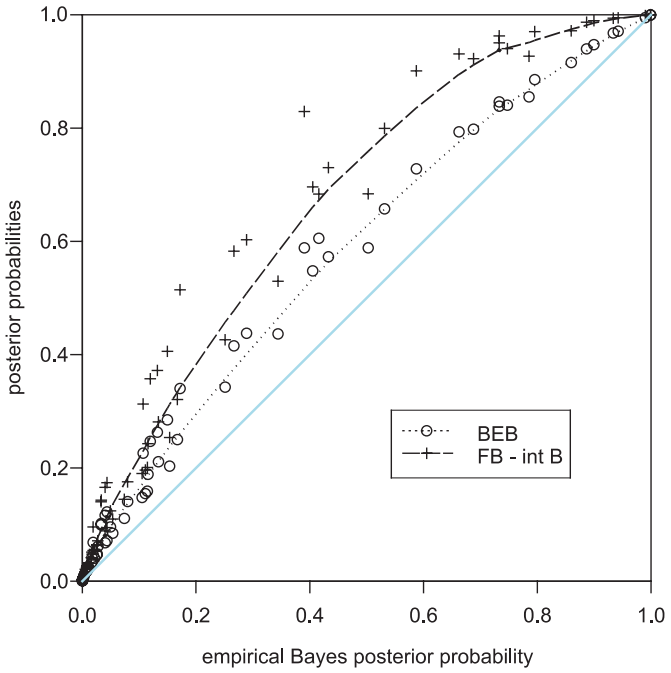
Power is related to the size of credible sets and therefore to posterior probabilities. When EB, BEB, and FB posterior probabilities are compared across the 91 sites of the HIV-1

Table 1. Sites identified to be under positive selection for the HIV-1 data set, before calibration. Lists are inferred by the naïve empirical Bayes (EB), Bayes empirical Bayes (BEB), and full Bayes (FB) methods.

	Positively selected sites							
	Frequentist		Full Bayes (FB)					
Analysis model	EB	BEB	meanIPP	medianIPP	PR_{95}	PMeanGO	PMedGO	PDGO
M2a (positive selection: 4 free parameters)	**28T 66E** 87V	**28T 66E** 87V	**28T 66E** 87V	**28T 66E 87V**	28T 66E	40 sites	37 sites	22 sites
M8a (beta & ω: 5 free parameters)	**28T 66E** 87V	26N **28T** 51I **66E** **87V**	22S 24E **26N** **28T 51I 66E** 68N **69N** 76E 83I **87V**	1V 22S 24E **26N** **28T** 39H **51I** **66E** 68N **69N** 76E **83I 87V**	26N **28T** 51I **66E** 69N 83I **87V**	38 sites	36 sites	25 sites

Note: meanIPP, the largest posterior mean of integrated posterior probabilities; medianIPP, the largest median of integrated posterior probabilities; PR_{95}, the percentile rank of integrated posterior probabilities; PMeanGO, the posterior mean greater than 1; PmedGO, the posterior median greater than 1; PDGO, the posterior density greater than 1. Cut-off thresholds = 0.95; (0.99 shown in bold). For PMeanGO, PMedGO, and PDGO, only the number of identified sites is provided.

Fig. 2. Probability–probability plot comparing empirical Bayes posterior probabilities (x-axis) with Bayes empirical Bayes (BEB) posterior probabilities (dotted line) and full Bayes (FB) mean integrated posterior probabilities meanIPP (dashed line) for the HIV-1 data set. Analyses assume model M8a. Solid line indicates major diagonal; dotted and dashed lines fitted through smoothing splines.



data set, EB probabilities tended to be smaller than BEB posterior probabilities, which were themselves smaller than FB posterior probabilities (Fig. 2). As a result, the FB approach tended to detect more sites under positive selection than BEB (Table 1). This suggests that FB is more powerful than BEB, but the difference might be due to higher false-positive rates. There appeared to be a nonlinear relationship between these posterior probabilities (Fig. 2), which is in contrast to the lack of relationship between EB and FB posterior probabilities found by Huelsenbeck and Dyer (2004). This difference is probably not due to their integrating over tree topologies; topologies have little impact on the inference procedure (Kosakovsky Pond and Frost 2005*b*; Yang et al. 2000). However, integrating over equilibrium codon frequencies, which adds 60 free parameters to the model, might cause some difficulties, especially with small data sets.

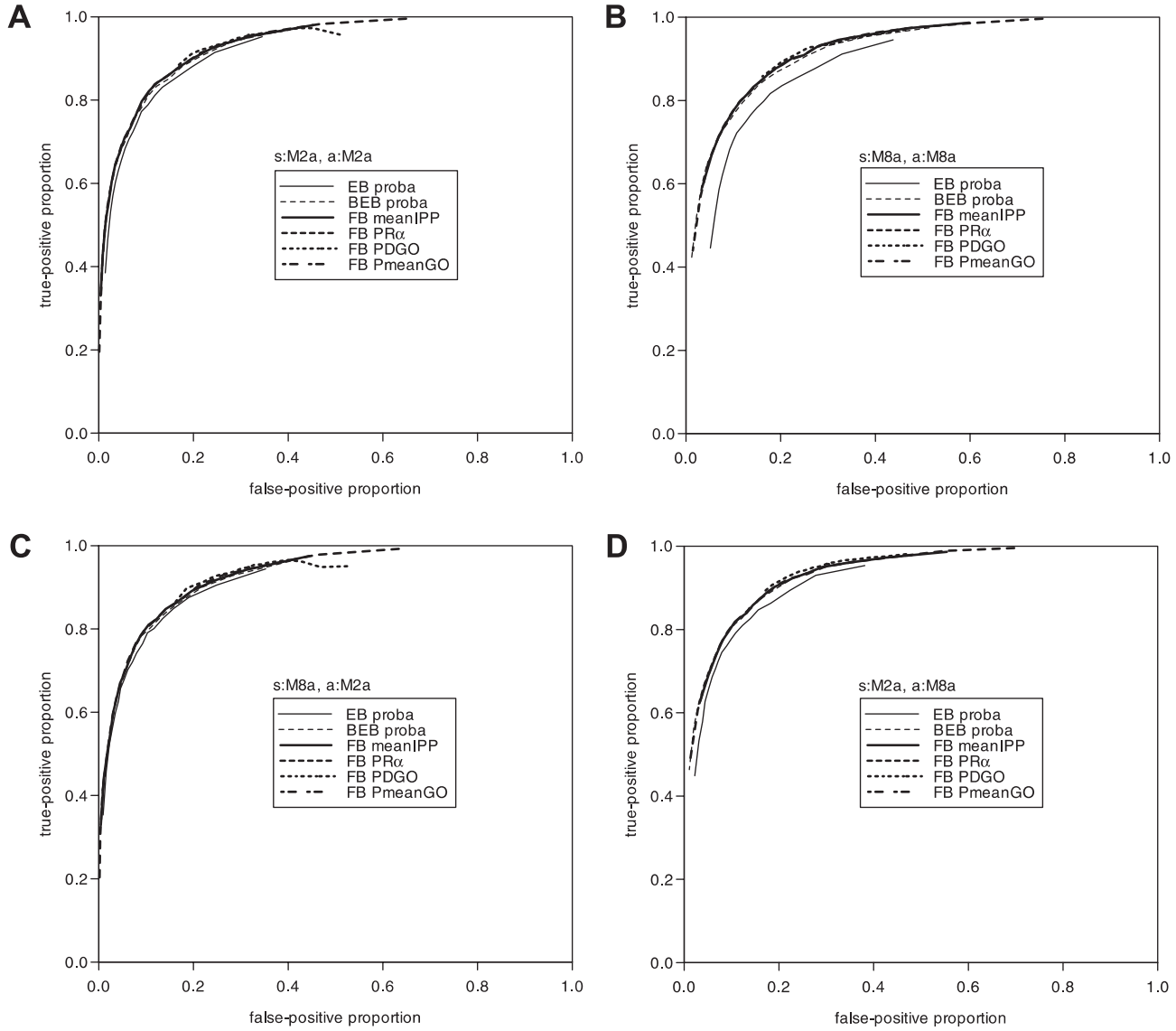
Although more sites were selected when integrating over uncertainties, the sites identified by both meanIPP (or medianIPP) and PR_{95} shared similar distributional features. They had either a point mass distribution on 1 (e.g., 28T or 66E) or a highly skewed distribution (e.g., 26N or 69N) (not shown). Huelsenbeck and Dyer (2004), however, observed a large degree of variability among posterior probabilities that sites were under positive selection. Part of the difference can be explained by our use of more stringent identification criteria. Whereas ω was >1 for all the sites identified using the criteria based on posterior probabilities, the reciprocal was not true (Table 1). Thirty-eight sites had integrated posterior mean ω values >1 (Table 1, PMeanGO), 25 had 95% of the posterior distribution above ω = 1 (Table 1, PDGO), and most had integrated posterior probabilities below the 95% cut-off threshold.

Integrating over nuisance parameters of both the random model describing among-site variation of ω and the substitution model appeared to increase the power to detect sites under positive selection. However, this result was obtained by comparing posterior probabilities on a real data set rather than on simulated data, and at a prespecified and identical threshold across the different criteria. These criteria are Bayesian, but because they treat nuisance parameters according to either frequentist (EB) or Bayesian (FB) practices, their respective thresholds might have different statistical implications. In addition, PR_{95} has 2 thresholds, making any direct comparison even more hazardous. As a result, a mere examination of Table 1 is likely to be misleading. A more statistically sound approach is required to make these criteria comparable.

Comparative analysis of the identification criteria

An ultimate goal when selecting a binary classifier is to maximize both its sensitivity (i.e., its ability to detect true positives) and its specificity (i.e., its ability to detect true negatives). This discrimination ability can be assessed by means of an ROC analysis. Originating from signal-detection theory, this analysis measures how well a receiver is able to detect a signal in the presence of noise, independent of the

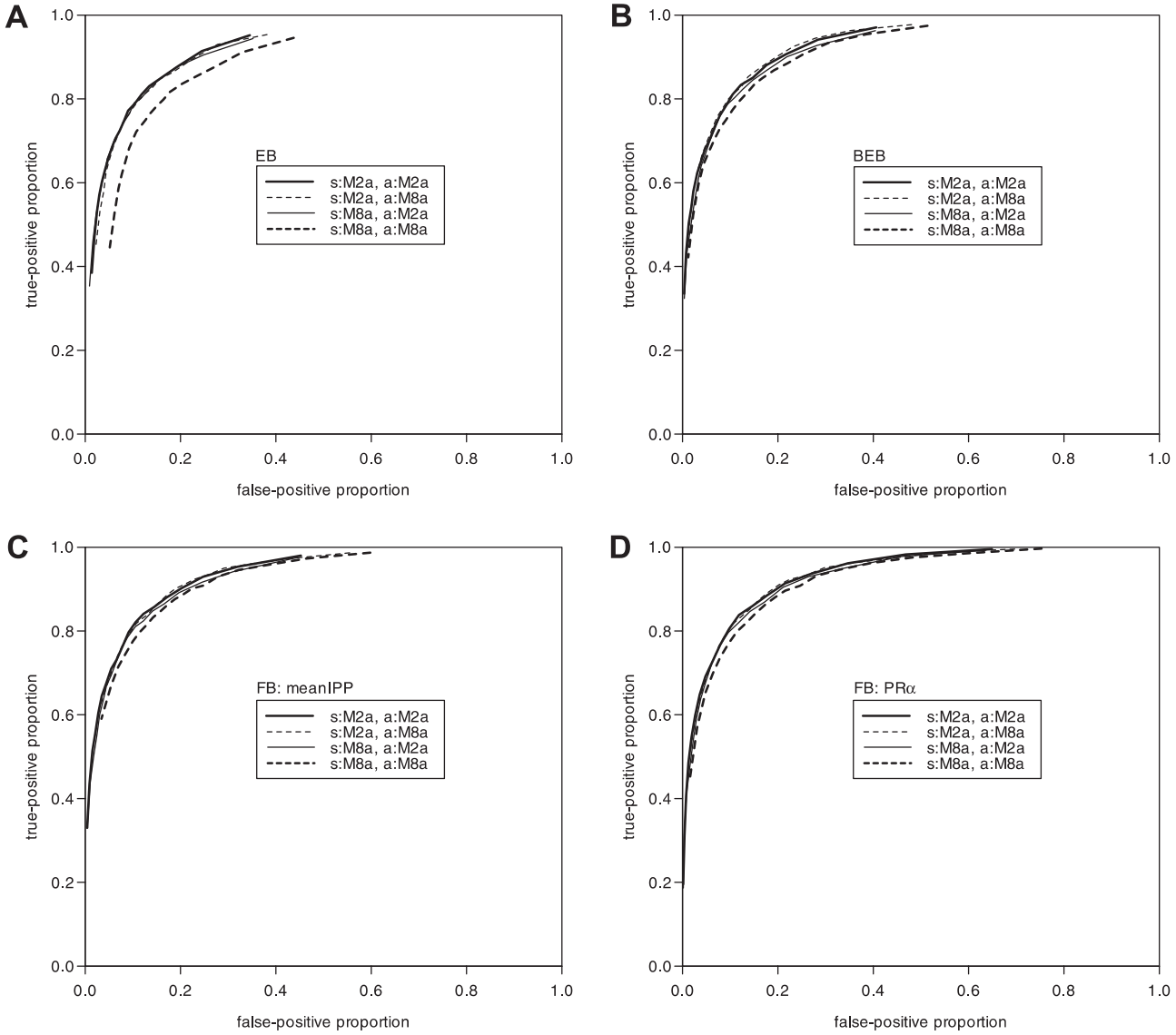
Fig. 3. Comparative predictive value of the different criteria. Receiver operating characteristic (ROC) curves of the criteria evaluated where the analysis model is correctly specified (A) in a simple scenario (s:M2a, a:M2a) and (B) in a more complicated scenario (s:M8a, a:M8a), and (C) where the analysis model is either underspecified (s:M8a, a:M2a) or overspecified (D) (s:M2a, a:M8a).



frequency of the event (here, positive selection), of the decision criterion, and of any prior assumption (e.g., Swets 1988). ROC analysis is summarized by an ROC curve, which is a graphical representation of the trade-off between true- and false-positive rates (1 – specificity) for every possible cut-off value. The more closely the ROC curve follows the left-hand border and the top border of the ROC space, the more discriminating the criterion is. Alternatively, a curve on the 45° diagonal (major diagonal) of the ROC space corresponds to a random decision process, with no discrimination. Figure 3 shows that all the criteria tested here performed better than a random decision process. However, differences exist, depending on whether nuisance parameters were integrated out or not and on the complexity of the models. Indeed, the EB criterion performed more poorly than any that integrate over nuisance parameters, especially when data were simulated under a complex model (M8a) and when analyzed under the same complex model (Fig. 3B). The Bayesian criteria under BEB and FB had very similar discriminating performances, irrespective of model specification. However, Fig. 4 shows that when a complex model is used to analyze the data mimicking the HIV-1 data set, discrimination is reduced, particularly under the empirical Bayes approach (Fig. 4A, contrast with Figs. 4C and 4D). It then appears preferable to analyze the data mimicking the small HIV-1 data set under a simple model, such as M2a, rather than a more complex model, such as M8a, irrespective of how the data were generated. This is likely due to the small number of sites (91) used in both the real-data analysis and the simulations.

If approaches integrating over nuisance parameters such as BEB and FB have comparable discrimination (i.e., true- and false-positive rates), how come the analysis of the HIV-1 data set identified different sites under different models (Table 1)? This discrepancy could result from the lack of calibration of the criteria (e.g., Andrews 1970). Calibration

Fig. 4. Effect of model specification on the predictive value of some criteria: (A) the empirical Bayes (EB); (B) the Bayes empirical Bayes (BEB); (C) the full Bayes (FB) meanIPP; and (D) the full Bayes percentile rank of integrated posterior probabilities at a given threshold α (PR_α) approaches.



simply amounts to finding a linear transformation to establish a correspondence between the criteria. Correspondence can be based on the actual probabilities of detecting a signal or on false-positive rates. Figure 2 suggests that such a transformation can be found, at least locally. However, calibration does not explain how, for example, EB detects 3% of the sites, whereas, at the other extreme, PMeanGO detects 44% at a given threshold (95%). That would mean that the threshold used by PMeanGO would have to be lowered to 0.95/(44/3) = 0.06, which, as a posterior probability, does not make much sense. The lack of calibration in Table 1 is then unlikely to explain the different number of sites identified, for example, by EB and FB. An alternative explanation is that the conflicting results in Table 1 suggest that FB approaches are less robust than empirical Bayes approaches. Because the specification of a prior distribution restricts the likelihood model, FB approaches are expected to be more sensitive to model specification, as suggested here by the discrepancy between the analysis of real and simulated data.

Practical considerations

Performing simulations to assess the predictive value of different codon models and site-identification criteria can be demanding. Yet, it made it possible here, on the small HIV-1 data set, to suggest that a complicated codon model, such as M8, could lead to higher false-positive rates than a simpler codon model, such as M2a (Fig. 4).

In actuality, this latter use of ROC curves is equivalent to comparing nonnested models, such as M2a, with M8a. Asymptotics for the likelihood-ratio test are complicated in this case (White 1982) and simpler approaches, such as the Akaike information criterion (AIC), are gaining popularity (Posada and Buckley 2004). Defined as $AIC_M = l + 2\ k$,

where l is the optimized log-likelihood under model M and k is the number of free parameters entering M, the criterion selects the model that is closest to the true or generating model (i.e., the model that minimizes AIC_M). Here, using the MLEs under models M2a and M8a (see also Yang et al. 2005), we obtained AIC_{M2a} = 2220.90 and AIC_{M8a} = 2222.78, respectively, so that the selected model is M2a. This is the model that was also selected by the ROC analysis. Minimizing the distance to the true process (AIC) and maximizing both sensitivity and specificity (ROC) are intuitively related objectives. A rigorous analysis of the relationship between AIC and ROC is missing at the moment. It is possible that using AIC to select a codon model that allows adaptive evolution prior to identifying sites might help reduce the false-positive rates caused by potential model misspecification when using Bayes approaches, such as BEB or FB.

Conclusions

Recent simulation studies have shown that, in terms of statistical performance, sitewise methods performed roughly as well as empirical Bayes methods (Kosakovsky Pond and Frost 2005*b*), and that empirical Bayes methods were outperformed by BEB approaches on small data sets (Yang et al. 2005). By transitivity, BEB approaches are expected to perform better than sitewise methods, at least on small data sets. Here, we showed that FB methods can perform better than empirical Bayes methods on small data sets (real and simulated data). However, we found that FB methods had only a marginal advantage over BEB on these small data sets (Fig. 3). This confirms the results of a recent study (Scheffler and Seoighe 2005) that also showed that FB methods perform even better when sequence divergence is small.

As FB methods are computationally demanding, we evaluated some simple methods of alleviating this burden and showed that fixing branch lengths to their MLEs did not affect the identification of sites under positive selection, at least on the HIV-1 data set used. Similar heuristic processes and approximations could prove particularly interesting when analyzing larger data sets. However, the FB approach remains computationally more demanding than BEB, and might be preferred over the latter only when uncertainty over parameters of the codon-substitution model itself is a source of concern, as can be the case in small data sets. The choice of which parameter(s) of the substitution model should be integrated over might be assessed by estimating corresponding maximum-likelihood confidence intervals.

In addition to better sensitivity and specificity than empirical Bayes approaches, the BEB and FB methods appeared in our simulations to be relatively insensitive to mild misspecification of the model used to account for rate heterogeneity among sites. This is in stark contrast to our analysis of empirical data, and suggests that Bayesian methods are less robust than empirical Bayes methods. As noted by Deely and Lindley (1981), the principal disadvantage of Bayesian methods is that they require the specification of prior distributions and hyperparameters, which generally impose some restrictions on the likelihood model. These restrictions can increase power when the model is correctly specified, but may cause adverse effects otherwise, to the extent that FB methods can be outperformed by empirical Bayes methods. For want of more realistic codon models, we suggest that a balance between the improved performance of BEB or FB and sensitivity to model specification might be reached using a model-selection criterion, such as AIC over nonnested codon models, which allow detection of adaptive evolution. Future work should focus on the relative performance of these different site-identification criteria when models are more severely misspecified than in this study, such as when recombination occurs (Anisimova et al. 2003) or when both nonsynonymous and synonymous rates vary among sites (Kosakovsky Pond and Frost 2005*b*). This latter point might be the most worrying from a theoretical point of view, but its biological meaning might also demand a clearer understanding and justification (Kosakovsky Pond and Frost 2005*a*).

Acknowledgements

I am grateful to Wendy Wong and Ziheng Yang for sharing a submitted manuscript. This work was funded by the Natural Sciences and Engineering Research Council of Canada (DG 311625) and by a startup fund from the University of Ottawa.

References

Andrews, D.F. 1970. Calibration and statistical inference. J. Am. Stat. Assoc. **65**: 1233–1242.

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

Anisimova, M., Nielsen, R., and Yang, Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics, **164**: 1229–1236.

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

Aris-Brosou, S., and Yang, Z. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst. Biol. **51**: 703–714.

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

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

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

Kosakovsky Pond, S.L., and Frost, S.D. 2005*a*. A simple hierarchical approach to modeling distributions of substitution rates. Mol. Biol. Evol. **22**: 223–234.

Kosakovsky Pond, S.L., and Frost, S.D. 2005*b*. Not so different after all: a comparison of methods for detecting amino-acid sites under selection. Mol. Biol. Evol. **22**: 1208–1222.

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

Leitner, T., Kumar, S., and Albert, J. 1997. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J. Virol. **71**: 4761–4770.

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

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

Posada, D., and Buckley, T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. **53**: 793–808.

Scheffler, K., and Seoighe, C. 2005. A Bayesian model comparison approach to inferring positive selection. Mol. Biol. Evol. **22**: 2531–2540.

Suzuki, Y., and 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.

Swanson, W.J., Nielsen, R., and Yang, Q. 2003. Pervasive adaptive evolution in Mammalian fertilization proteins. Mol. Biol. Evol. **20**: 18–20.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. Science, **240**: 1285–1293.

White, H. 1982. Maximum likelihood estimation of misspecified models. Econometrica, **50**: 1–26.

Wong, W.S., Yang, Z., Goldman, N., and 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. 2001. Adaptive molecular evolution. *In* Handbook of statistical genetics. *Edited by* D.J. Balding, M. Bishop, and C. Cannings. Wiley, London. pp. 327–350.

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

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