Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection

Maria Anisimova,† Joseph P. Bielawski,* and Ziheng Yang**

*Department of Biology, Galton Laboratory and †Center for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London

Bayes prediction quantifies uncertainty by assigning posterior probabilities. It was used to identify amino acids in a protein under recurrent diversifying selection indicated by higher nonsynonymous (d_N) than synonymous (d_S) substitution rates or by $\omega = d_N/d_S > 1$. Parameters were estimated by maximum likelihood under a codon substitution model that assumed several classes of sites with different ω ratios. The Bayes theorem was used to calculate the posterior probabilities of each site falling into these site classes. Here, we evaluate the performance of Bayes prediction of amino acids under positive selection by computer simulation. We measured the accuracy by the proportion of predicted sites that were truly under selection and the power by the proportion of true positively selected sites that were predicted by the method. The accuracy was slightly better for longer sequences, whereas the power was largely unaffected by the increase in sequence length. Both accuracy and power were higher for medium or highly diverged sequences than for similar sequences. We found that accuracy and power were unacceptably low when data contained only a few highly similar sequences. However, sampling a large number of lineages improved the performance substantially. Even for very similar sequences, accuracy and power can be high if over 100 taxa are used in the analysis. We make the following recommendations: (1) prediction of positive selection sites is not feasible for a few closely related sequences; (2) using a large number of lineages is the best way to improve the accuracy and power of the prediction; and (3) multiple models of heterogeneous selective pressures among sites should be applied in real data analysis.

Introduction

One of the ways to study adaptive molecular evolution is by identifying the amino acid sites where the nonsynonymous substitution rate (d_N) exceeds the synonymous rate (d_S). Only when nonsynonymous mutations offer a selective advantage are they fixed at a higher rate than synonymous mutations. The difference between these two rates is measured by the ratio $\omega = d_N/d_S$, with $\omega > 1$ indicating evolution by positive selection. (We use the terms ‘‘adaptive evolution,’’ ‘‘positive’’ and ‘‘diversifying’’ selection interchangeably, although the methods we examine here are mainly suitable for detecting recurrent diversifying selection and may lack the power of detecting directional selection or episodic adaptive evolution.) Models assuming the same ω ratio for all sites only detect positive selection if the average $\omega > 1$ (Yang and Bielawski 2000). However, most amino acid sites are functionally conserved (with ω close to 0), and only a few are responsible for molecular adaptation (Golding and Dean 1998). Moreover, these few sites might not be clustered in a sequence because sites that are far apart in a primary sequence can be clustered in the three-dimensional structure. Thus, even a sliding window analysis might not detect positive Darwinian selection in many genes (e.g., Endo, Ikeo, and Gojobori 1996).

A number of authors proposed alternative methods that account for variable selective pressures across sites. Fitch et al. (1997) and Suzuki and Gojobori (1999) used the parsimony method to reconstruct sequences in the extinct ancestors and counted the changes along the tree at each site to identify sites at which there was an excess of nonsynonymous substitutions. Nielsen and Yang (1998) and Yang et al. (2000) implemented codon substitution models of heterogeneous ω ratios among sites in the maximum likelihood (ML) framework and used the Bayes method to predict sites under positive selection. ML and Bayes methods are major statistical estimation methods and are widely used in molecular phylogenetics (Huelsenbeck and Rannala 1997; Lewis 2001; Whelan, Lio, and Goldman 2001). Although ML models are computationally intense for large numbers of lineages, they do not rely on ancestral reconstruction and can easily accommodate known features of sequence evolution, such as the transition-transversion rate bias and the codon usage bias. The ML approach uses the likelihood ratio test (LRT) to compare two nested models: a null model, which does not account for sites with $\omega > 1$, and an alternative model that does. A gene is considered to be under positive selection if (1) the LRT is significant, and (2) at least one of the ML estimates of $\omega > 1$. When the ML parameter estimates indicate the presence of sites under positive selection, the empirical Bayes approach can be used to predict them (Nielsen and Yang 1998; Yang et al. 2000). One computes the posterior probability that a site belongs to each ω class of the model, given the data at that site. Sites with high posterior probabilities of belonging to a class with $\omega > 1$ are likely to be evolving by positive selection.

The codon models were successfully applied to detect positive selection in a number of genes. For example, Bishop, Dean, and Mitchell-Olds (2000) detected strong adaptive pressure in the cell wall–attacking enzyme chitinase and mapped positively selected residues on a three-dimensional structure of the class I chitinase.

Key words: Bayes inference, likelihood, nonsynonymous-synonymous rate ratio, positive selection, posterior probability.

Address for correspondence and reprints: Ziheng Yang, Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom.
E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 19(6):950–958. 2002

An excess of amino acid replacements in the active cleft of the enzyme indicated that class I chitinase evolves in response to pathogenic variation. This supported a hypothesis of coevolutionary ‘‘arms race’’ between plants and their pathogens. The same approach demonstrated that evasion of the immune system by viruses occurs by diversifying selection, e.g., across the HIV-1 *nef* gene (Zanotto et al. 1999) and the capsid gene of the foot-and-mouth disease virus (Fares et al. 2001; Haydon et al. 2001). Remarkably, there was a correlation between the known cytotoxic T-lymphocytes (CTL) epitopes responsible for antigenic determination and the predicted positive selection sites.

A better understanding of adaptive evolution, however, also requires an understanding of how the adaptive changes affect phenotypes. The Bayes method can identify the few sites responsible for adaptive change, providing the initial information required to understand the changes in the form and function of proteins over evolutionary time. Specific structural and functional hypotheses can be formulated if we know which sites in an ancestral protein evolve by positive selection (e.g., Adey et al. 1994; Chandrasekharan et al. 1996; Dean and Golding 1997; Yang, Swanson, and Vacquier 2000). Site-directed mutagenesis could then be used to conduct biochemical testing of such hypotheses (Chang, Kazmi, and Sakmar 2001).

Although the ML method appeared successful in real data analysis, its performance is not well understood. Anisimova, Bielawski, and Yang (2001) used computer simulation to examine the performance of the LRT in detecting adaptive evolution; they found that the use of the χ^2 distribution made the LRT conservative; however, the test was powerful when enough lineages were analyzed. In this study we examine the performance of Bayes prediction of positive selection sites. We simulate data under heterogeneous ω models and evaluate the performance of the method under different sequence lengths, sequence divergences, and numbers of taxa. We explore conditions under which Bayes posterior probabilities give a reliable measure of uncertainty.

Theory and Methods

Codon Substitution Models of Heterogeneous ω Ratios Among Sites

The following models for the distribution of ω among sites are used in this paper either for simulating or for analyzing the data: M0 (one-ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta), and M8 (beta&ω) (Yang et al. 2000). M0 (one-ratio) assumes the same ω ratio for all sites. M1 (neutral) assumes two site classes with $\omega_0 = 0$ (conserved sites) and $\omega_1 = 1$ (neutral sites). M2 (selection) adds an extra class to M1 with an ω_2 estimated from the data. Under model M3 (discrete), sites in the sequence are drawn from K discrete classes with the ω ratios $\omega_0, \omega_1, \ldots, \omega_{K-1}$ taken in proportions $p_0, p_1, \ldots, p_{K-1}$. In this paper $K = 3$. Model M7 (beta) assumes a beta distribution $B(p,q)$, with parameters p and q. Because the beta distribution is limited within the interval (0,1), it provides a flexible null hypothesis for testing positive selection. Model M8 (beta&ω) adds another site class with a constant ω. In PAML (Yang 1997) the beta distribution is approximated by 10 equal-probability discrete categories, with the median ω value for each category calculated using the parameters p and q (Yang et al. 2000).

Positive selection, that is, the presence of sites with $\omega > 1$, can be tested by LRTs by comparing models M0 (one-ratio) with M3 (discrete), M1 (neutral) with M2 (selection), and M7 (beta) and M8 (beta&ω). The performance of two LRTs, which compare M0 with M3 and M7 with M8, was examined by Anisimova, Bielawski, and Yang (2001).

Bayes Prediction of Sites Under Positive Selection

We estimate the parameters of the codon model by ML and use the Bayes theorem to infer to which ω class each site is most likely to belong (Nielsen and Yang 1998; Yang et al. 2000). Suppose a model of heterogeneous ω ratios assumes K classes, with the proportions and ω ratios given as

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

Proportions p_i are the prior probabilities for site classes, $i = 0, 1, \ldots, K - 1$. The posterior probability that site h with data x_h belongs to class i is

$$P(\omega_i \mid x_h) = \frac{P(x_h \mid \omega_i)p_i}{P(x_h)} = \frac{P(x_h \mid \omega_i)p_i}{\sum_{j=0}^{K-1} P(x_h \mid \omega_j)p_j}. \tag{2}$$

Here, $P(x_h|\omega_i)$ is the probability of the data at site h, given that it belongs to site class i. In the implementation of models by Nielsen and Yang (1998) and Yang et al. (2000), parameters such as ω_i and p_i in equation (1) are estimated by ML, and the ML estimates are used to calculate the posterior probabilities in equation (2). This is known as the empirical Bayes approach. If a site has a high posterior probability of coming from a class with ω estimated to be >1, the site is likely to be under positive selection. A conservative approach can be taken to predict sites under positive selection by requiring this posterior probability to exceed a cut-off value such as $P = 0.95$ or 0.99. Thus, a site with a high probability (say 0.9) of being under positive selection is not considered to be under positive selection if a stringent cutoff is applied (say $P = 0.95$).

Simulations

All results were based on 100 simulated replicates. All data sets were simulated under models M3 (discrete) and M8 (beta&ω) and included a small fraction of sites evolving under positive selection (table 1). ML estimates of parameters from 17 vertebrate β-globin genes (Yang et al. 2000) formed the basis of most simulations (table 1). Simulated data sets varied in the number of taxa T, sequence length in the number of codons L_c, sequence divergence S, and selective pressure, here ω_2 for M3 and ω for M8 (table 1). S was measured by tree length, i.e., the expected number of nucleotide substi-

Table 1
Parameter Values Used in Simulations

Model System	Simulation Model	Parameters in the ω Distribution	Sequence Length, L_c	Number of Taxa, T	Tree Length, S
β-Globin genes of vertebrates	Discrete (M3)	$p_0 = 0.386$, $p_1 = 0.535$, **$p_2 = 0.079$**, $\omega_0 = 0.018$, $\omega_1 = 0.304$, **$\omega_2 = 1.691$**	100, 500	6	0.11, 1.1, 11
	Discrete (M3)	As above	100, 500	17	0.38, 2.11, 16.88
	Discrete (M3)	As above except **$\omega_2 = 4.739$**	100, 500	6	0.11, 1.1, 11
	Beta&ω (M8)	$p_0 = 0.943$, $p = 0.572$, $q = 2.172$, **$p_1 = 0.057$, $\omega = 2.081$**	500	6	1.1
	Beta&ω (M8)	As above	500	17	8.44
Hemagglutinin genes of human flu virus	Discrete (M3)	$p_0 = 0.470$, $p_1 = 0.470$, **$p_2 = 0.060$**, $\omega_0 = 0.018$, $\omega_1 = 0.304$, **$\omega_2 = 3.143$**	329	98	1.3

NOTE.—Simulation parameters representing positive selection are indicated in bold. The beta distribution in M8 was discretized into 10 equal probability classes with ω ratios 0.002, 0.015, 0.038, 0.070, 0.112, 0.165, 0.234, 0.323, 0.446, and 0.651.

tutions per codon along the tree. We used three values for S (table 1) and referred to them as low, medium, and high sequence divergences, respectively. We also simulated data sets using the ML parameter estimates obtained from 98 hemagglutinin gene sequences of the human influenza type A virus (GenBank accession numbers AF180564–AF180666; except for duplicates AF180572, AF180577, AF180596, AF180636, and a highly divergent outlier AF180666; table 1). These sequences constitute a subset of the data analyzed by Bush et al. (1999) and Yang (2000).

Three unrooted trees were used in simulation: (1) an artificial 6-taxon tree, (2) a 17-taxon tree with branch lengths estimated from the vertebrate β-globin genes, and (3) a 98-taxon tree with branch lengths estimated from the influenza A hemagglutinin gene. Trees (1) and (2) were used previously to evaluate the properties of the LRT to detect positive selection (see fig. 1 of Anisimova, Bielawski, and Yang 2001). Tree (3) was similar in shape to the influenza trees presented by Bush et al. (1999) and Yang (2000). Different levels of sequence divergence were generated by multiplying all branch lengths by a scale factor to achieve the desired tree length (S). Data were simulated under models M3 and M8, using the program evolver in the PAML package (Yang 1997). An ω ratio and an ancestral codon state at the root of the tree are randomly drawn from the given multinomial distributions. Then the program ‘‘evolves’’ each site along the branches of the tree independently, according to the Markov process of codon substitution (Goldman and Yang 1994; Muse and Gaut 1994). Sites evolving by positive selection (with $\omega > 1$) are listed in a file and are later compared with the predicted sites.

Analysis

Simulated data were analyzed by ML under models M3 (discrete) or M8 (beta&ω), using the codeml program in the PAML package (Yang 1997). The branch lengths in the tree were estimated by ML iteration, together with the parameters in the substitution models. Positive selection sites predicted by codeml (eq. 2) were compared with the true sites from evolver. We did not perform the LRTs before Bayes prediction of the selected sites.

Accuracy of Bayes prediction was measured by the probability that a site predicted to be under positive selection was truly under positive selection. Let $N_{\bullet +}$ denote the number of sites predicted to be under positive selection (table 2). Here, the first subscript refers to classifications by evolver, and the second subscript refers to prediction by codeml, with ‘‘•’’ meaning all sites, ‘‘+’’ meaning sites with $\omega > 1$, and ‘‘−’’ meaning sites with $\omega < 1$. For example, N_{++} denotes the number of correctly predicted positive selection sites. Accuracy was estimated by averaging the proportion $N_{++}/N_{\bullet +}$ over the replicates. Replicates in which positive selection was not detected ($N_{\bullet +} = 0$) were ignored in the calculation. Note that prediction of sites under positive selection by codeml depends on a cut-off probability P.

The power of Bayes prediction was measured by the probability that a site truly under positive selection (from evolver) was predicted to be under positive selection by codeml. Thus, for each replicate the power was calculated as $N_{++}/N_{+\bullet}$, where $N_{+\bullet}$ is the number of all sites that are truly under positive selection (table 2). Here, we include replicates in which no positive selection was predicted ($N_{++} = N_{\bullet +} = 0$). However, we ignored replicates in which all estimates of ω were >1 ($N_{\bullet +} = L_c$ and $N_{++} = N_{+\bullet}$); this happened on rare occasions for very similar sequences if no or very few synonymous substitutions occurred. In such cases, ω estimates are highly unreliable.

Although it is convenient to think of false positives and false negatives for the prediction of positive selection sites, the Bayes method is not based on formal hypothesis testing. Our use of the terms accuracy and power in this paper is informal, and they are not equal to (1 − type I error rate) and (1 − type II error rate) in hypothesis testing. One could measure accuracy by $N_{--}/N_{-\bullet}$, so that both accuracy and power ($N_{++}/N_{+\bullet}$) are conditional on the truth (classification of sites by evolver). However, prediction of conserved sites or sites not under positive selection is easy and biologically not so interesting.

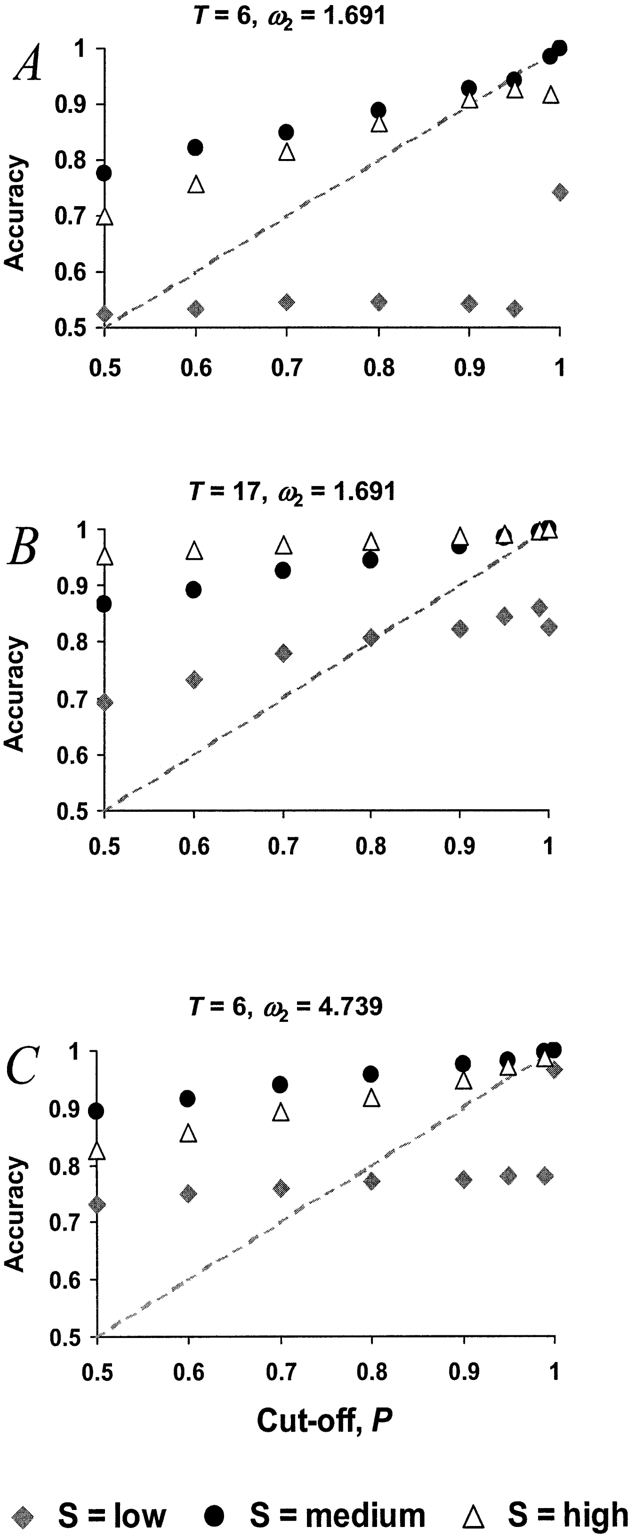

FIG. 1.—Accuracy of the Bayes prediction under M3 (discrete) for sequences of 500 codons. Simulated data varied in the number of lineages ($T = 6, 17$), sequence divergence (S, see table 1), and strength of selective pressure ($\omega_2 = 1.691, 4.739$).

Table 2
Partition of Sites in a Sequence Used in the Estimation of Accuracy and Power of Bayes Prediction

		PREDICTED SITES (codeml)		
		+	−	Total
True sites (evolver)	+	N_{++}	N_{+-}	$N_{+\cdot}$
	−	N_{-+}	N_{--}	$N_{-\cdot}$
	Total	$N_{\cdot+}$	$N_{\cdot-}$	L_c

NOTE.—The first subscript refers to classifications by evolver and the second subscript refers to prediction by codeml, with "+" meaning sites under positive selection with $\omega > 1$ and "−" meaning sites not under positive selection. Accuracy is calculated as $N_{++}/N_{\cdot+}$ and power as $N_{++}/N_{+\cdot}$. Note that the Bayes prediction by codeml depends on a cutoff value P.

Results

Accuracy of Bayes Prediction

For all parameter combinations accuracy increased only very slightly with the increase in sequence length from $L_c = 100$ to 500 codons. Hence, the results for $L_c = 100$ codons are not shown. Figure 1 presents simulation results concerning the accuracy of Bayes site prediction under the discrete model (M3) for sequences of 500 codons. Each point shows the estimated probability that a site predicted by codeml at a cut-off P is truly under positive selection, i.e., in the list from evolver. Bayes prediction is trustable only if accuracy is higher than the cut-off value P.

Accuracy was always much higher for sequences of medium and high sequence divergences than for those at low sequence divergence (fig. 1). The worst case was observed for data sets of six sequences at low divergence ($S = 0.11$), where accuracy was much lower than the cut-off values (fig. 1*A*). Note that tree length $S = 0.11$ means highly similar sequences, with 0.024 nucleotide substitutions per nonsynonymous site and 0.078 nucleotide substitutions per synonymous site along the tree. In such data sets, only about 10% of the sites (codons) showed any variability, and even fewer sites had two or more changes. Even at medium and high sequence divergences ($S = 1.1$, $S = 11$), accuracy was lower than the cut-off P for $P > 0.8$ ($L_c = 100$) or $P > 0.9$ ($L_c = 500$; fig. 1*A*). Also note that high divergence ($S = 11$) means very divergent sequences, with 2.4 nonsynonymous substitutions per nonsynonymous site and 7.8 synonymous substitutions per synonymous site on the small tree consisting of six taxa. The results of figure 1 suggest that Bayes prediction is tolerant of multiple substitutions at the same site.

Accuracy was improved by increasing the number of taxa to 17 (compare fig. 1*A* with *B*). However, for similar sequences ($S = 0.38$) accuracy remained well below the cut-off P for $P > 0.7$ ($N = 100$; results not shown) or $P > 0.8$ ($N = 500$; fig. 1*B*). For medium and high sequence divergences ($S = 2.11$, $S = 16.88$), accuracy was always higher than the cut-off P (fig. 1*B*). We note that the best accuracy was achieved at medium divergence in the small tree (fig. 1*A*) and at high divergence in the large tree (fig. 1*B*). Although sequence divergences in trees of different sizes are not directly comparable, we expect large trees to be more tolerant of

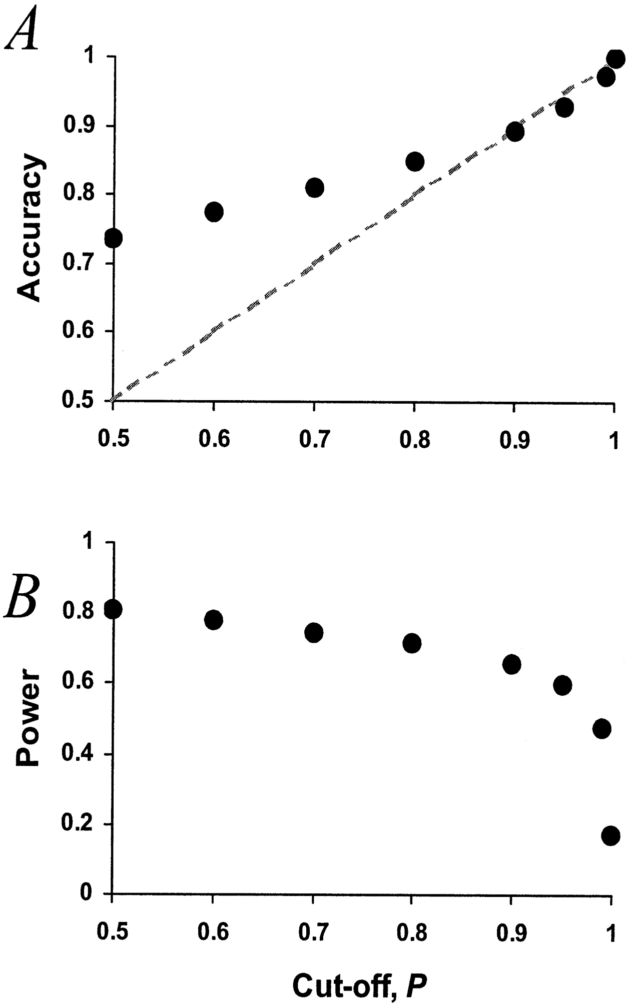

FIG. 2.—Accuracy (*A*) and power (*B*) of Bayes site prediction in data sets of $T = 98$ sequences, using parameter estimates obtained from the hemagglutinin gene of human influenza type A virus ($N = 329$, $S = 1.3$; see table 1).

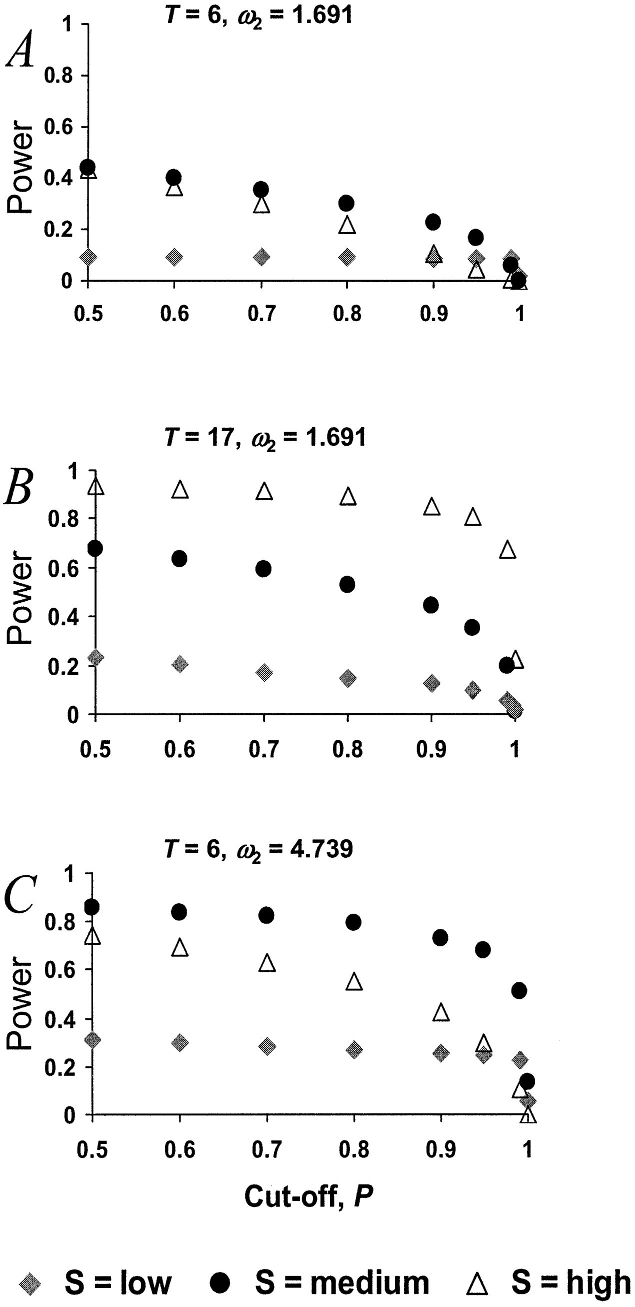

FIG. 3.—Power of the Bayes prediction under M3 (discrete) for sequences of 500 codons. See caption of figure 1.

multiple substitutions. We also analyzed data in which the strength of positive selection was higher ($\omega_2 = 4.74$) and observed a substantial improvement in accuracy at all levels of sequence divergence (compare fig. 1*C* with *A*). Although in reality one has no control over the level of positive selection pressure on a gene, Bayes prediction is likely to be more accurate in the presence of strong positive selection.

We simulated data sets of 98 lineages using the ML parameter estimates from the influenza A hemagglutinin gene under model M3 (discrete) (table 1). The results are presented in figure 2. The tree length $S = 1.3$ means an average branch length of $S/(2T - 3) \approx 0.007$ nucleotide substitutions per codon, so that the sequence divergence is even lower than the low divergences in the 6-taxon ($S = 0.11$) and 17-taxon ($S = 0.38$) trees in previous experiments. However, accuracy was much higher in the 98-taxon data sets (compare fig. 2*A* with fig. 1*A*, *B*, and low *S*). Accuracy was higher than the cut-off values when $P < 0.90$ and only slightly lower than the cut-off values when $P > 0.90$ (fig. 2*A*). For example, accuracy was ≈0.93 at $P = 0.95$ and ≈0.97 at $P = 0.99$ (fig. 2*A*).

Power of Bayes Prediction

Power to predict positive selection sites was strongly dependent on both the level of sequence divergence and the number of taxa but was largely unaffected by sequence length. Figure 3 presents power of Bayes site prediction under the discrete model (M3) for sequences of 500 codons. Each point in figure 3 shows the estimated probability that a site truly under positive selection (from evolver) is predicted by codeml at the cut-off *P*. In general, the higher the cut-off *P*, the fewer are the sites predicted to be under positive selection. Thus, the power decreases with the increase in the cutoff.

Regardless of the number of taxa, power was much higher for medium and high sequence divergences than for low divergence. For example, for 17-taxon data sets with $L_c = 500$, the power at the 0.95 cutoff was 0.10, 0.35, and 0.81 for low, medium, and high sequence divergences, respectively (fig. 3*B*). Increasing the number of taxa from 6 to 17 yielded substantial increases in power, especially at medium and high divergences (fig. 3*A* and *B*). For very similar sequences, power improved substantially by increasing the number of taxa to 98 (fig. 2*B*). For example, at the 0.95 cutoff, power was 0.08 (fig. 3*A*, low *S*), 0.14 (fig. 3*B*, low *S*), and 0.60 (fig. 2*B*) for 6-, 17-, and 98-taxon data sets, respectively.

We observed much higher power for data sets simulated with increased positive selection pressure ($\omega_2 = 4.74$; compare fig. 3*C* with *A*). Thus, one is more likely to predict positive selection sites if strong positive selection acts on the gene.

The effect of the tree shape reflected in relative branch lengths was not examined in this paper but is expected to be similar to that of the sequence divergence. A well-balanced tree, with an even distribution of changes over its branches, will achieve optimal information content. Highly biased trees might harbor too many changes along the long branches and too few changes along the short branches, leading to lack of information and low power in the detection method.

Discussion

Sampling Errors of ML Estimates of Parameters and Their Effect on the Accuracy of Bayes Site Prediction

Because the empirical Bayes method uses parameter estimates in the prior distribution, the accuracy of the prediction is affected by errors in the estimates of the parameters. Indeed, if the parameters are correct, the posterior probabilities calculated from equation (2) will be the exact probabilities that the site is under positive selection. Thus, the low accuracy of Bayes site prediction in certain parameter combinations reported in this paper (fig. 1) is the result of sampling errors in the ML parameter estimation. We expect Bayes prediction (eq. 2) to be much more sensitive to parameters in the ω distribution, such as the proportions and ω ratios in model M3 (discrete), than to other parameters such as branch lengths of the tree. Models of variable ωs among sites, such as M3 (discrete) and M8 (beta&ω) used in this paper, are known as finite mixture distributions. Parameters in such mixture distributions are notoriously difficult to estimate partly because of strong correlations among parameter estimates (Johnson, Kotz, and Kemp 1993, pp. 309–322). For example, a small proportion of sites under strong positive selection (with a larger ω) might fit the data almost equally well as does a slightly larger proportion of sites under less strong selection (with a lower ω), and typical data sets will not contain sufficient information to distinguish between the two. Results of figure 1 suggest that this problem is serious when the data contain only a few highly similar sequences, where the information content is very low.

The alternative in this case is the hierarchical Bayes approach (or full Bayes approach), which accounts for uncertainties in the parameters in the prior distribution by assigning and integrating over a hyper-prior distribution for parameters in the prior. This does not increase the power of the analysis but has the effect of adding noise into the model so that the posterior distribution will become more spread out, reducing the confidence in the prediction. The computation can be achieved by the Markov chain Monte Carlo algorithms but is expected to be much more expensive than for the ML procedure implemented in PAML. For small data sets, where the use of the full Bayes approach is most rewarding, the computation seems feasible and well worth pursuing.

In the mean time, we advise caution on Bayes prediction of sites under selection when the data contain only a few highly similar sequences. Collecting more sequences seems to be the most effective strategy in improving the accuracy and power of the analysis.

A further difficulty is caused by the fact that ω is a continuous variable and is best described by a continuous distribution, but we had to use a discrete distribution for computational reasons (Yang et al. 2000). As an example, if the three ω ratios under M3 are 0.9, 1.0, and 1.1, it will be extremely difficult to assign sites into these classes using the Bayes or any other approach. Consequently, one should be very cautious when drawing a conclusion about positive selection when the estimated ω is only slightly >1.

Robustness Analysis and Testing Selection with Multiple Models

To evaluate the robustness of Bayes prediction of sites under positive selection and to explore the performance of different models, we simulated data sets under models M3 and M8 and analyzed them using three different models: M2 (selection), M3 (discrete), and M8 (beta&ω). Both 6-taxon and 17-taxon trees were used in the simulation, and we present results obtained from the 17-taxon data sets in figure 4. Note that the parameters for M3 and M8 were obtained from the same real data set (table 1). When model M3 was used in both simulation and analysis (diamonds in fig. 4*B* and *D*), the accuracy and the power of Bayes site prediction were similar to those when M8 was used for both simulation and analysis (circles in fig. 4*A* and *C*). Regardless of the simulation model, a lower accuracy and a higher power were achieved when data were analyzed under M3 rather than under M8. The difference is because of the fact that M3 predicted almost 30% more sites than M8 did. The accuracy and the power of site prediction under M2 (selection) were essentially the same as under M8 (beta&ω) (fig. 4). It should be noted, however, that M2 was very conservative for six-taxon data sets, i.e., accuracy was very high, but power was very poor (results not shown).

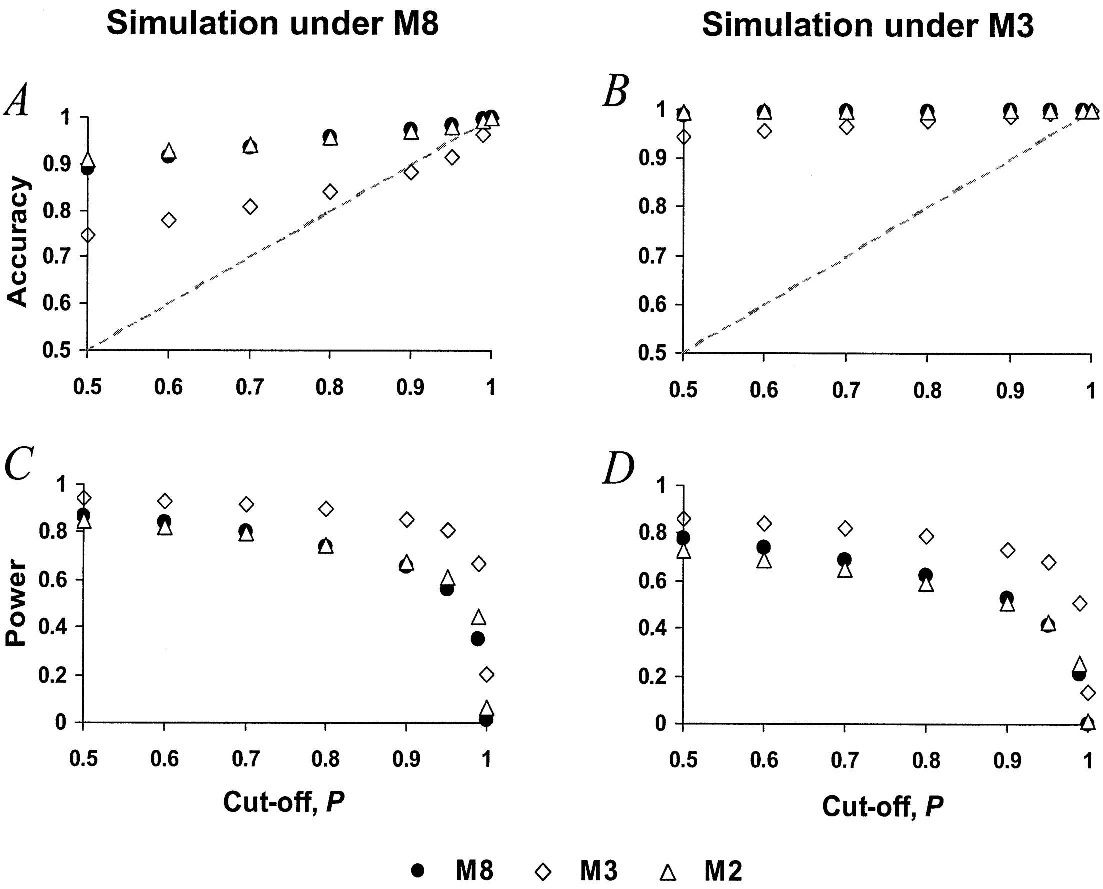

FIG. 4.—Bayes prediction of selected sites using different models: M2 (selection), M3 (discrete), and M8 (beta&ω). Data were simulated under M3 and M8 with $S = 8.44$, $N = 500$, $T = 17$, and $\omega = 2.081$ (see table 1).

The differences among the models appear to be mainly the result of their different formulations. In this sense, Bayes prediction of selected sites appears insensitive to the model assumed in generating the data. The different properties of the models have become apparent in the analysis of real data sets as well as in simulation data (fig. 4, see also Yang et al. 2000; Anisimova, Bielawski, and Yang 2001). For example, M8 (beta&ω) makes a strict distinction between sites evolving with $\omega < 1$ and those with $\omega > 1$ and is more stringent than M3 in detecting selective pressure on a gene (discrete). However, M8 is computationally intense and is known to have multiple local optima in some data sets. To avoid being trapped at a local optimum, it is important to run M8 at least twice, once with initial $\omega > 1$ and once with $\omega < 1$, and results corresponding to the highest likelihood value should be used (see PAML manual; Yang 1997). Although M2 is less flexible than M8, its use of a neutral site class with $\omega_1 = 1$ means that sites undergoing neutral evolution or under weak positive selection will be lumped into that class, reducing false positives. However, M2 can be very conservative. The corresponding null model M1 (neutral) does not account for sites with $0 < \omega < 1$, and the additional class in M2 may be forced to account for such sites, with positively selected sites lumped into the class with $\omega_1 = 1$, leading to the failure to detect positive selection (Yang et al. 2000).

Model M3 is not as precise in distinguishing positive selection sites as M8 or M2. When the strength of positive selection is low, and there is a large fraction of neutral sites, M3 can yield a large number of incorrectly predicted sites. Anisimova, Bielawski, and Yang (2001) also showed that violations of model assumptions could lead to a high rate of false positives for the LRT in detecting positive selection. However, M3 has some advantages. First, M3, with only three site classes, appears to fit any data set as well as any other model (e.g., Yang et al. 2000). Second, it does not appear to have multiple local optima and is computationally fast.

Because each model has advantages and disadvantages, we suggest the use of multiple models in real data analysis. Yang et al. (2000) recommended the following models for real data analysis: M0 (one-ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta), and M8 (beta&ω).

Overall Recommendations Concerning the Likelihood Method to Detect Positive Selection

The likelihood models implemented by Nielsen and Yang (1998) and Yang et al. (2000) are designed to perform the following tasks, in order of increasing difficulty: (1) test for the presence of sites under positive selection ($\omega > 1$) by LRT, comparing two nested models, (2) estimation of the proportion of positive selection sites and the strength of positive selection by ML estimation of parameters in the ω distribution, and (3) identification of amino acid sites under positive selection by Bayes prediction.

The LRT seems quite reliable even in small data sets with only a few similar sequences, according to the simulation study of Anisimova, Bielawski, and Yang (2001). Although use of the χ^2 approximation makes the test conservative, the test appeared quite powerful. We expect some data sets to contain insufficient information for reliable identification of sites under positive selection, but it is still interesting to know whether such sites exist at all. In such data sets, the LRT, which uses combined evidence from all sites in the sequence, should be more powerful than methods that test positive selection at each site by counting substitutions (e.g., Suzuki and Gojobori 1999).

Estimation of parameters in the ω distribution is a much more difficult matter, mainly because of the strong correlation among parameters in the ω distribution. The different models of variable ωs among sites and different parameters of the same model correspond to the different ways of lumping sites into classes, and for a given data set, there can be many almost equally good ways of lumping sites.

Lastly, Bayes prediction of sites under selection is most difficult, and many sequences are required to accumulate synonymous and nonsynonymous changes at individual sites, which are the source of information for inferring the underlying ω ratio at each site. Furthermore, random and systematic errors in parameter estimates will affect the accuracy of Bayes prediction. Also, note that the accuracy and power measures used in this paper apply to one site and not to the entire sequence, and that it is almost certain that some sites in the sequence will be incorrectly identified, because so many inferences are made in one analysis.

Given these considerations, it is remarkable that the application of these methods to real data has generated biologically highly sensible results (e.g., Zanotto et al. 1999; Bishop, Dean, and Mitchell-Olds 2000; Swanson et al. 2001). We note that our simulations examined conditions where Bayes site prediction will be most difficult, i.e., small data sets with a few highly similar sequences. To date, most ML-based studies of positive selection in nonviral organisms sampled more than 17 lineages, which represented relatively divergent sequences (e.g., Bishop, Dean, and Mitchell-Olds 2000; Yang, Swanson, and Vacquier 2000; Peek et al. 2001). Bayes prediction in such cases should be more reliable.

Based on our simulations, we make the following generalizations. (1) Prediction of positively selected sites is unreliable when sequences are very similar, and the number of lineages is small (e.g., $S \leq 0.11$ or $T \leq 6$). (2) Increasing the number of lineages is the most effective way to improve accuracy and power. Accurate prediction is possible for data sets comprising very similar sequences if a very large number of lineages have been sequenced. (3) Multiple models should be used in real data analysis to ensure the robustness of the results.

Acknowledgments

We thank Willie Swanson and two anonymous referees for their comments. We are grateful to A. Philip Dawid for discussion on our measures of accuracy and power. This study was funded by grants from the Biotechnology and Biological Sciences Research Council and the Human Frontier Science Program to Z.Y. and by a Medical Research Council studentship to M.A.

LITERATURE CITED

Adey, N. B., T. O. Tollefsbol, A. B. Sparks, M. H. Edgell, and C. A. Hutchison III. 1994. Molecular resurrection of an extinct ancestral promoter for mouse L1. Proc. Natl. Acad. Sci. USA **91**:1569–1573.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test to detect adaptive molecular evolution. Mol. Biol. Evol. **18**:1585–1592.

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

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. **16**:1457–1465.

Chandrasekharan, U. M., S. Sanker, M. J. Glynias, S. S. Karnik, and A. Husain. 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. Science **271**: 502–505.

Chang, B. S. W., M. A. Kazmi, and T. P. Sakmar. 2001. Synthetic gene technology: applications to ancestral gene reconstruction and structure-function studies of receptors. Methods Enzymol. **343**:274–294.

Dean, A. M., and G. B. Golding. 1997. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc. Natl. Acad. Sci. USA **94**:3104–3109.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

Fares, M. A., A. Moya, C. Escarmis, E. Baranowski, E. Domingo, and E. Barrio. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol. Biol. Evol. **18**:10–21.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94**:7712–7718.

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

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

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

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

Johnson, N. L., S. Kotz, and A. W. Kemp. 1993. Univariate discrete distributions, Vol. 1. Wiley, New York.

Lewis, P. O. 2001. Phylogenetic systematics turns over a new leaf. TREE **16**:30–37.

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

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

Peek, A. S., V. Souza, L. E. Eguiarte, and B. S. Gaut. 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (fimA) from Escherichia coli. J. Mol. Evol. **52**:193–204.

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

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

Whelan, S., P. Lio, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. **17**:262–272.

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

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

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. TREE **15**:496–503.

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

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

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

Adam Eyre-Walker, reviewing editor

Accepted February 15, 2002