J Mol Evol (2013) 76:280–294
DOI 10.1007/s00239-013-9557-0

The Site-Wise Log-Likelihood Score is a Good Predictor of Genes under Positive Selection

Huai-Chun Wang · Edward Susko · Andrew J. Roger

Received: 26 November 2012 / Accepted: 20 March 2013 / Published online: 18 April 2013
© Springer Science+Business Media New York 2013

Abstract The strength and direction of selection on the identity of an amino acid residue in a protein is typically measured by the ratio of the rate of non-synonymous substitutions to the rate of synonymous substitutions. In attempting to predict positively selected sites from amino acid alignments, we made the unexpected observation that the site likelihood of an alignment column for a given tree tends to be negatively correlated with the posterior probability that site is in the positive selection class under widely-used codon models. This is likely because positively selected sites tend to be more variable and display more "radical" amino acid changes; both of these features are expected to result in low site log-likelihoods. We explored the efficacy of using the site log-likelihood (SLL) score as a predictor for positive selection. Through simulation we show that a SLL-based test has a low false positive rate and comparable power as the codon models. In one case where the simulated data violated the assumption that synonymous substitution rates were constant across the sites, the codon models were not able to detect positive selection in the data while the SLL test did. We applied the new method to ten empirical datasets and found that it made similar predictions as the codon models in eight of them. For the *tax* gene dataset the SLL test seemed to produce more reasonable results. The SLL methods are a valuable complement to codon models, especially for some cases where the assumptions of codon models are likely violated.

Keywords Positive selection · Codon models · Multiple tests · Synonymous rate · Nonsynonymous rate · Maximum likelihood

Electronic supplementary material The online version of this article (doi:10.1007/s00239-013-9557-0) contains supplementary material, which is available to authorized users.

H.-C. Wang (✉) · E. Susko
Department of Mathematics and Statistics, Dalhousie University, Halifax, NS B3H 4R2, Canada
e-mail: hcwang@mathstat.dal.ca

H.-C. Wang · A. J. Roger
Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 4R2, Canada

H.-C. Wang · E. Susko · A. J. Roger
Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS B3H 4R2, Canada

Introduction

The inference of codons under positive selection from a protein-coding sequence alignment has traditionally involved estimating the ratio of non-synonymous substitutions rates (dN) versus synonymous substitutions rates (dS) at a site. The ratio of dN/dS, commonly denoted as ω being greater than, equal to, or less than 1 indicates the gene is under adaptive (positive), neutral mutation or purifying (negative) selection, respectively. Many statistical methods have been developed to infer positive selection from protein-coding DNA sequences and identify these sites by comparing the dS and dN along the codon sequences (Yang 2006). The most sophisticated of these are the random effects codon models, which assume a discretized distribution of the ω's across the codon sites and infer the most probable ω class at this site given this distribution under a maximum likelihood/empirical Bayes framework (Anisimova and Kosiol 2009).

There are more than a dozen models that employ these kinds of ω distributions and many of these have been

implemented in the phylogenetic analysis by maximum likelihood (PAML) package, a commonly used software tool for detecting positive selection (Yang 2007). One of the assumptions of all of these models is that the synonymous substitution rates are constant across sites and only the variation of the non-synonymous substitution rates among sites is modeled through variation in ω. However, it should be noted that this is not a realistic assumption for many protein-coding genes. For example it is well-known that in bacteria and some eukaryotes such as yeast and *Drosophila* synonymous codon usage in highly expressed genes is biased toward those codons that have the most abundant corresponding tRNA in the cell to maximize growth rates (Ikemura 1985; Sharp and Li 1987), indicating that synonymous mutations in these organisms are not neutral and that the rate of their fixation varies over loci. In addition, Hurst and Pal (2001) give an example involving BRCA1 where variation in the dN/dS ratio over sites is largely due to variation in the synonymous rates. Such cases can create difficulties for codon models as it was recently shown in Rubinstein et al. (2011) that variation in the rates of synonymous substitutions can inflate false positive rates (FP).

Furthermore, the estimation of the ratio of dN/dS requires that the sequences in comparison cannot be too close or too divergent (Yang and dos Reis 2011). When the sequences are too divergent, synonymous substitutions reach saturation and dS may be underestimated, which may inflate the ω ratio. High sequence divergence can cause other problems, such as differences in codon usage between the species or errors in the alignment, which may lead to false prediction of positive selection (Yang and dos Reis 2011; Privman et al. 2012; Jordan and Goldman 2012). When the sequences are too similar, the numbers of synonymous and non-synonymous substitutions will be insufficient for reliable estimation of the model parameters and thus reduce the power for detecting positive selection (Nozawa et al. 2009; Nickel et al. 2008), or conversely it may lead to an elevated false positive prediction rate in extremely sparse data sets (Suzuki and Nei 2004; but see also Yang et al. 2005). Furthermore, codon models simply cannot be validly applied to some kinds of data sets, such as those that include a mixture of sequences from organisms (or organelles) with differing genetic codes (Knight et al. 2001).

In light of these issues, several workers have proposed methods for detecting positive selection independent of ω (Tang and Wu 2006; Hanada et al. 2007; Shapiro and Alm 2009; Zhou et al. 2010). For example, the study by Shapiro and Alm proposed using the ratio of substitutions in slow- versus fast-evolving sites (S/F) to quantify variation in natural selection on specific branches compared to the typical pattern of selective constraint observed across species in the phylogeny. This S/F ratio relies on empirical definitions of "slow" and "fast" sites rather than predefined synonymous and non-synonymous sites so that the new test can work on amino acid sequences alone or on anciently divergent codon sequences for which ω-based methods are not applicable due to synonymous substitutions being saturated with multiple substitutions.

It has been shown that sites under positive selection are often unusually variable on the amino acid level and frequently display functionally/structurally "radical" amino acid changes (Hughes and Hughes 1993; Zhang 2000; Pupko et al. 2003; Hanada et al. 2007). We noticed that these sites also tend to have much lower site log-likelihood scores than sites that are classified as being under purifying selection ($\omega < 1$) or neutral evolution ($\omega = 1$). Indeed, we find that there is a strong negative correlation between amino acid site log-likelihood scores (lnL) and two indicators for selective strengths in simulated data: the estimated ω values of the codon sites and the Bayes posterior probabilities of sites being in a positive selection class. This property allows the development of methods based on the site lnL score in amino acid alignments to predict the protein families under positive selection as well as identify positively selected sites. Using simulations we show these methods have very low FP and reasonable power in predicting positive selection at the gene level if the sequences are not too divergent or too similar. We also find that in simulations based on codon models that are different from the codon models used in estimation, the site-likelihood based methods have power to detect positive selection in contrast to some standard codon models that have little or no power under these conditions. Finally, we test the new methods on ten empirical datasets and show that identical predictions to those of codon models are obtained for eight data sets on the gene level and most of the same sites are identified as positively selected. However, for one of the datasets previously suggested to confound inferences of positive selection using codon site models (Suzuki and Nei 2004), the site-wise likelihood prediction method appeared to give more reasonable predictions than standard codon models.

Methods

Site-Wise Amino Acid Log-Likelihood (SLL) and Thresholds for Declaring a Site Positively Selected and Gene-Wise Testing of Positive Selection

For a dataset of codon sequences, simulated or empirical, we first fit parameters on a phylogenetic tree and obtained the amino acid site log-likelihoods under a standard amino acid substitution model such as the WAG + F + Γ model

for the translated amino acid sequences. We then use simulation under an appropriate null model to determine a site likelihood threshold below which we declare a site positively selected. Fixing the tree topology as inferred from the amino acid sequences, we estimate the codon data under a M8A codon model (Swanson et al. 2003) with equal codon frequencies (every codon has a frequency of 1/61) to get estimates of the model parameters, using Codeml in the PAML package (Yang 2007). Parameters include edge-lengths, κ, the differing ω parameters for the different site classes and associated weights or probabilities of a site being in these classes. Since the M8A model assumes ω at a site comes from a mixture of a discrete version of the two-parameter beta-distribution ($\omega < 1$) and a point mass at $\omega = 1$, no positive selection is allowed. Using the M8A tree, estimated parameters and codon frequencies we simulate a dataset of 20,000 codon sites under M8A and analyze it under the WAG + F + Γ model for the translated amino acid sequences to get a null distribution of amino acid site log-likelihoods under no positive selection. The bottom 5th percentile of the 20,000 amino acid site log-likelihoods was used as a critical value, i.e., if a site in the original data has a log-likelihood lower than this value, then it is identified as a putative "positive" site.

Site likelihoods are most naturally used to locate sites that are under positive selection. However, in many applications (e.g., where many gene families are being examined at once), a test of whether a gene, as a whole, is under selection is also desired. Given any site-wise test with a false-positive rate of α, the expected proportion of sites declared positive is α if the gene is not under selection, and this expected proportion is larger than α if the gene is under selection. Thus, if the site-wise tests are independent, a one-sided binomial test provides a general way of constructing a gene-wise test from a site-wise test. (In our case the site-wise tests are not independent because all sites are used to obtain the parameter estimates used to obtain the site likelihoods, but this is a relatively small departure from independence.) Specifically, defining θ to be the proportion of the putative positively selected sites found in the original data, if the site-wise test is an α-level test, the gene is declared to be under positive selection if θ greater than θ_c at a certain α level.

$$\theta_c = \alpha + Z_{\alpha'}\sqrt{\frac{\alpha(1-\alpha)}{n}} \quad (1)$$

where n is the number of the codon sites in the original data. Here α and α' are typically set to 0.05 giving $Z_{\alpha'} = 1.645$. The gene-wise test then has a FP of α'.

A subtly different issue in declaring a site positively selected is to adjust for the multiple tests or selection bias: although it is relatively rare that the p value for a randomly selected site, not under selection, will be <0.05, it is not unlikely that the smaller p values among n independent p values will be <0.05. A general approach to correcting for this selection bias was provided by Benjamini and Hochberg (1995). The procedure declares sites corresponding to the k smallest p values to be positively selected when the rth of these is $<r\alpha/n$, $r = 1,\ldots, k$ but the $k + 1$st is larger than $(k + 1)\alpha/n$; no site is declared positively selected if the smallest p value is larger than α/n. The Benjamini–Hochberg criterion has the property that the overall probability is α that at least one site is falsely declared positive. It complements the uncorrected site-wise test which has false positive probability α for any particular site but is likely to give at least one false positive for a large gene.

Simes (1986) describes a variation of the Benjamini and Hochberg criterion intended as an overall test rather than a multiple comparisons adjustment for tests at sites. It has been used previously to create gene-wise tests from site-wise ones in Wong et al. (2004). The procedure declares the gene to be under selection if the rth smallest p-value is smaller than $r\alpha/n$. Since the Benjamini–Hochberg procedure will not declare any sites positively selected if the smallest p-value is not smaller than α/n, the Simes test, while similar, may declare a gene to show evidence of selection even though no sites are found significant. Such a situation may arise, for instance, if no single site is very strongly selected for. As a gene-wise test, the Simes test complements the binomial test, as it is good at detecting a small subset of sites are under selection but requires that the signal for selection be strong. By contrast, the binomial test can detect that weak selection is taking place if it occurs at a sufficient number of sites. As the two tests differ in their relative strengths at detecting different types of gene-wise selection patterns, gene-wise predictions made under the two methods can be combined to increase the power of the tests, which we refer collectively as a SLL test.

Prediction of Genes under Positive Selection: Type I Error Rates and the Power

To investigate the Type 1 error (false positive) rates of the SLL methods, we used INDELIBLE (Fletcher and Yang 2009) under the codon M0 model to simulate 200 datasets each with 300 codons under a neutral model of evolution ($\omega = 1$ for all sites) for an equal codon frequency (all 1/61), under a symmetric bifurcating tree of 32 tips with the same branch lengths (all 0.1). These settings were first used in Suzuki and Nei (2002) who claimed they were difficult for the codon models because they induce high rates of false positives. The same tree and setting were also used in subsequent reports for comparing the FP of the codon models in detecting positive selection (Wong et al. 2004; Massingham and Goldman 2005; Pond and Frost 2005). For each

simulated dataset, we used the SLL approach to determine the number of wrongly assigned positive sites and compare their proportion (θ) with a critical θ_c (=0.071 for $n = 300$ and $\alpha = 0.05$ based on Eq. 1) to decide if the null model of neutral evolution was rejected and the dataset was falsely inferred as positively selected. Simultaneously we applied the Simes test to the simulated datasets to find FP under its criterion. As a comparison, for the 200 simulated datasets we used two commonly used likelihood ratio tests (LRT) to determine FP from the codon models: M8 versus M7 (Yang et al. 2000a, b) and M8 versus M8A (Swanson et al. 2003). For the M8 + M7 pair test we compared the LRT statistic (twice of the log-likelihood difference between the two models) with $\chi^2_{2,\,0.05} = 5.991$. For the M8 + M8A pair test, there are two χ^2 thresholds, one is based on a $^{1/2}\chi^2_0 +^{1/2} \chi^2_1$ mixture with a critical value of 2.71 (Swanson et al. 2003) and another is the more conservative $\chi^2_{1\ ,0.05}$ with a critical value of 3.84 (Wong et al. 2004). From the M8 estimation results, the numbers of the falsely assigned positive sites with the naïve empirical Bayes (NEB) and the Bayes empirical Bayes (BEB) probabilities (Yang et al. 2005), both >95 %, were recorded separately.

After showing the type 1 error rates for both gene-wise and site-wise tests are close to the nominal 0.05, we further studied the power of the SLL tests. The simulations for the power analyses were based on an abalone sperm lysin gene tree (Fig. 1 in Yang et al. 2000b). The original abalone lysin data had 25 taxa and 135 codon sites and was shown to be under positive selection with various codon sites models (Yang et al. 2000b). For example, analyzed under the discrete M3 model with F3 × 4 codon frequencies, we obtained the following model parameters: the transition to transversion ratio parameter $\kappa = 1.5761$; selection class 1 ($\omega 1$) = 0.08519 with weight $p1 = 0.32902$; selection class 2 ($\omega 2$) = 0. 9112 with weight $p2 = 0.40231$; and selection class 3 ($\omega 3$) = 3.06543 with weight $p3 = 0.26868$. These parameter estimates were used for the following simulations and the three ω's corresponded to strong purifying selection, weak purifying or nearly-neutral case and strong positive selection respectively, which made it an ideal case to study the power of relevant methods for detecting positive selection. Based on the codon frequencies under the F3 × 4 model, κ and the lysin tree we simulated the following six cases of positive selection:

Case 1: Simulations employed the M3 model with the same three omegas (ω) and their weight parameters (p) estimated as above.

Case 2: Simulations employed the M3 model with the ω and p parameters as above, but the branch lengths of the tree were increased by 10-fold. This setting was used to evaluate the performance of the methods on data that were potentially saturated with multiple substitutions.

Case 3: Simulations employed the M3 model with the ω and p parameters as above but the branch lengths of the tree were decreased by 10-fold. This setting generates data sets with little variation among sequences.

Case 4: Simulations employed the M3 model with a very "weak" positive selection regime of: $\omega 1 = 0.08519$, $p1 = 0.32902$; $\omega 2 = 0.\ 9112$, $p2 = 0.621$; $\omega 3 = 1.5$, $p3 = 0.05$. Here both $\omega 3$ and its weight were smaller than in Case 1. All other parameters were as described in the first case.

Case 5: Simulations employed the M3 model with a relatively "weak" positive selection regime of: $\omega 1 = 0.08519$, $p1 = 0.32902$; $\omega 2 = 0.\ 9112$, $p2 = 0.40231$; $\omega 3 = 1.5$, $p3 = 0.26868$. The difference between this case and Case 4 is that the proportion of the $\omega 3$ is much bigger in this case.

Case 6: Simulations generated three separate datasets each under a different M0 model with the following ω's: 0.08519; 0. 9112 and 3.06543, respectively, and the numbers of the sites in the three sets followed the same proportions of the three classes as in Case 1. The three datasets were then concatenated into one dataset. This situation, unlike the foregoing scenarios, allows the rate of synonymous substitution to be uncoupled and different across the three partitions of the dataset. Specifically, the expected numbers of synonymous substitutions per unit edge length for the simulated datasets of the three ω classes are 0.79, 0.26, and 0.09, respectively; the corresponding expected numbers of non-synonymous substitutions per unit edge length are 0.21, 0.74, and 0.91, respectively. This mimics the situations where the synonymous substitutions may not be strictly neutral and differ in rates across loci due to factors such as translational selection on highly expressed genes (e.g., Ikemura 1985).

For each of the six simulation scenarios, 100 replicates of simulated data with 200 codon sites were generated with INDELIBLE. The SLL tests were applied to the datasets to determine the number of the sets predicted to be positively selected and the predicted positive sites were recorded. For comparison, two pairs of LRTs (M8 + M7 and M8 + M8A) were used to determine the power of the standard codon models. For the analyses under the codon models, Codeml (PAML version 4.2b; Yang 2007) was used with the F3 × 4 codon frequency model and fixed the tree topology as the lysin tree used in generating the data. For the SLL method the codon sequence data were translated into amino acids and analyzed with Codeml under a standard amino acid model with four discrete Gamma rates (WAG + F + Γ) and fixing the tree topology as the lysin tree and allowing branch lengths optimized. Since Codeml did not directly give the amino acid log-likelihood at each

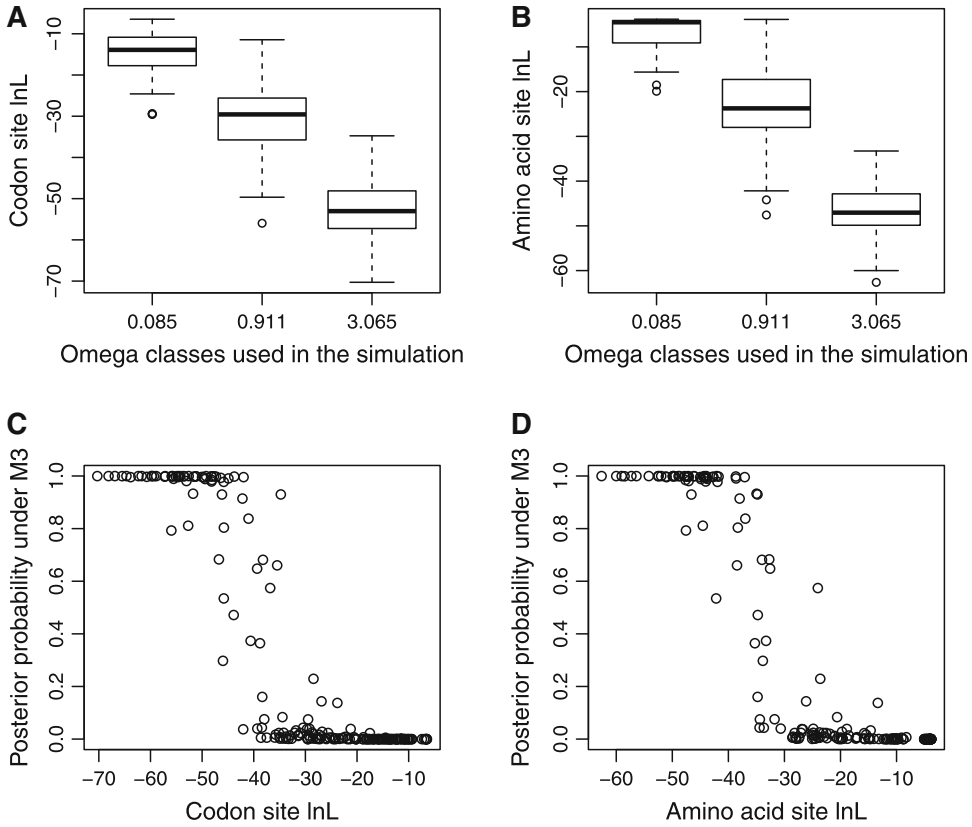

Fig. 1 **a** and **b**: six *box plots* showing the distributions, for the three ω classes, of site log-likelihoods from a dataset simulated under M3, with (**a**) being the codon site log-likelihoods estimated under M3 and (**b**) being the amino acid site log-likelihoods estimated under WAG + F + Γ for the translated amino acid data. **c** and **d** Two *scatter plots* showing the correlations between the posterior probabilities of the sites being in a positive selection class estimated under M3 and the site log-likelihoods, with (**c**) being the codon site log-likelihoods estimated under M3 and (**d**) being the amino acid site log-likelihoods estimated under WAG + F + Γ for the translated amino acid data

site but listed the log-likelihoods of the site patterns, we wrote a script to get the site-wise lnL's based on the site patterns.

Prediction of Sites under Positive Selection

For the SLL method, a site is considered to be under positive selection if its amino acid log-likelihood is smaller than the 5th percentile of the null log likelihood distribution for the 20,000 amino acid sites simulated under M8A. For codon models allowing a positive selection class (e.g., M3 and M8), positive sites are predicted based on the estimated empirical Bayesian posterior probabilities that a site belongs to a positive selection class ($\omega > 1$). Typically a site with a posterior probability of >0.5 (or more stringently 0.95) in an $\omega > 1$ class is assumed positively selected (Yang et al. 2000a, b). The posterior probabilities calculated under M3 are NEB that treat the estimated ω as fixed for each site and do not account for sampling errors in the maximum likelihood estimations of the model parameters. This may cause unreliable posterior probability calculations (Yang et al. 2005). The M8 model calculates, in addition to the NEB probabilities, the BEB of the sites, a generally more reliable method for determining sites under positive selection (Yang et al. 2005). For each simulated dataset in the six simulation cases, we predicted positively selected sites under the SLL test and under the M3 model for the NEB probabilities and the M8 model for both NEB and BEB probabilities (the BEB probabilities under M3 are not available for the PAML version we used).

To fairly compare the power of different methods for the same set of simulation parameters, their respective false-positive rates should be same. This poses a problem when comparing the SLL test with the codon models in detecting positively selected sites. The *p*-values of the SLL tests are based on a null distribution of strict neutrality ($\omega = 1$) that is the hardest case to distinguish from positive selection and thus is an upper bound on the *p*-value that one might obtain under the null hypothesis. Codon models produce Bayesian posterior probabilities and the rule that declares positives when posteriors are larger than 95 % need not have a 5 % FP (Massingham and Goldman 2005). The receiver operator characteristic (ROC) analysis avoids the difficulties of comparing the proportions of correctly predicted positive

sites for methods with different FP. This analysis, presented as a ROC curve, plots the true positive rates (TP) against FP (Green and Swets 1966) thus allowing comparison of TP at any given fixed FP. For the six cases in the power analyses simulated with INDELIBLE, we knew which sites were true positives as INDELIBLE assigned each site to a rate class corresponding to a given ω class. ROC curves were plotted for the following four predictors: the SLL; the NEB posterior probabilities under M3; the NEB probabilities under M8; and the BEB probabilities under M8. As the FP rates take on a discrete set of values that depends on the number of sites under selection which varies from data set to data set, the TP rates cannot be directly summed to get a set of mean TP rates for plotting an average ROC curve for each simulation case. To obtain the TP rates at the same set of FP rates for each data set in a simulation case, we applied a smooth spline function available as part of the R statistical programing environment (R Development Core Team, 2008). Average ROC curves over 100 data replicates are reported in each case.

Empirical Case Studies

We analyzed ten empirical datasets previously studied for positive selection, including the abalone sperm lysin gene (Yang et al. 2000a), the *tax* gene of a human T cell lymphotropic virus (Suzuki and Nei 2004), *Drosophila* alcohol dehydrogenase (ADH), Flavivirus E-glycoprotein, Flavivirus NS-5, the vertebrate β–globin, Japanese encephalitis *env* and three HIV-1 genes (*pol*, *vif* and *env* V3 region). The last eight datasets were among the ten datasets previously analyzed in Yang et al. (2000a, b) and downloaded from http://abacus.gene.ucl.ac.uk/ziheng/data.html; the other two datasets were not available from the web source. We used the SLL method to analyze the ten datasets with the trees previously built by the original authors with the branch lengths optimized. We also reanalyzed the data using three codon models (M7, M8, and M8A), although they were previously tested with various codon models (except M8A) in the original publications.

Results

Correlation Between Site Likelihood Score and Selective Strength

To explore the correlation between the site likelihood score and the selective strength (measured as ω), we simulated a dataset of 200 codon sites under the M3 model with three ω classes ($\omega = 0.0852$, 0.9112, and 3.06543, respectively) based on a 25-taxa lysin tree. Figure 1 top panel shows six box plots of the site log-likelihoods for the three classes of sites estimated under M3 (Fig. 1a) and the corresponding amino acid data analyzed under WAG + F + Γ (Fig. 1b). It indicates, for both the codon data and amino acid data, that sites under strong purifying selection (the first group: $\omega = 0.0852$) have the highest site likelihoods; sites under strong positive selection (the third group: $\omega = 3.06543$) have the lowest site likelihoods; sites under weak purifying selection (the second group: $\omega = 0.9112$) tend to have site likelihoods in between the two other classes. Regardless of whether amino acid or codon site likelihoods were considered, two-group t tests for a difference in mean site log likelihoods were very significant for any pairwise comparison of the ω classes ($p < 2.2 \times 10^{-16}$). Similar patterns as in Fig. 1a and b emerged when the codon site log-likelihoods were estimated under the M0 and M8 models and the amino acid site log-likelihoods were estimated under the equal rates (WAG + F) model (data not shown).

The bottom panel of Fig. 1 shows two scatter plots of the NEB probabilities estimated under M3 against the codon site log-likelihoods (Fig. 1c) and the amino acid site log-likelihoods under discrete Gamma model (Fig. 1d) for this data set. The two figures are very similar and present curves that look like the letter Z, as the posterior probabilities are bounded by 0 and 1, whereas site log-likelihoods are only bounded at 0. It is clear that sites with the lowest lnL have the highest posterior probabilities of positive selection and those with the highest lnL have the lowest posterior probabilities. Similar curves appeared when the posterior probabilities were plotted against the codon site log-likelihoods estimated under M0, M8 or the amino acid site log-likelihoods estimated under the equal rates model (figures not shown).

Amino acid and codon site likelihoods may thus be used as predictors of positive selection. In the following, we focus on the amino acid site log-likelihood scores estimated under the standard amino acid model (e.g., WAG + F + Γ) for testing positive selection and investigate its statistical properties (type I error rates and the power). Given the very similar behaviors for codon and amino acid site lnL's shown in Fig. 1, the results may apply to a codon model-based site likelihood predictor as well.

Type I Errors of Tests for Positive Selection

Of the 200 datasets of 300 codon sites simulated under a 32 taxa tree and a model of neutral evolution ($\omega = 1$ for all sites), none were predicted to be positive under the binomial test (the critical threshold $\theta_c = 0.071$ based on Eq. 1), while the Simes test predicted 13 sets being positive (6.5 %). By comparison, for the same 200 simulated datasets, the M8 + M7 test predicted five sets to be positive (type 1 error rate = 2.5 %), while the M8 + M8A test surprisingly predicted 39 sets (19.5 %) to be positive under

the χ^2 mixture criterion (see the Method) and 17 sets (8.5 %) to be positive under the χ^2 with 1degree of freedom criterion. Because the mixture χ^2 distribution criterion had such a high FP it was not used in further analyses.

The number of wrong positive sites, among all 60,000 sites from the 200 data sets (at cutoff levels 0.95 and 0.99) and related measures are listed in Table 1. The results show that the SLL test had on average a FP <3 %; the NEB prediction of the M8 model had a very high FP (over 40 % on average and reached to 100 % in many datasets) and the BEB prediction had a very small rate under the 0.95 and 0.99 cutoffs. The high FP of the NEB predictions highlights the importance of conducting prescreening for positive datasets before using the NEB approach to predict positively selected sites (Anisimova et al. 2002; Massingham and Goldman 2005).

Power of the Gene Level Tests

For each of the six cases of positive selection (see the Methods) 100 data sets each of 200 codon sites were simulated and analyzed with the SLL and codon LRT tests. For 200 sites, the critical θ_c at the standard $\alpha = 0.05$ level is 0.075 according to Eq. 1. However, since the actual observed FP for site-wise predictions was only 3 %, this θ_c calculated under the nominal α may be too conservative for determining a gene-wise prediction of positive selection. To adjust for the low FP in the site-wise predictions, we used $\alpha = 0.03$ in calculating a critical value for the binomial test and $\theta_c = 0.03 + 1.645\sqrt{\frac{0.03*(1-0.03)}{200}} = 0.05$. In Table 2, we list the power of the binomial test based on the two θ_c values (at $\alpha = 0.05$ and 0.03) as well as that of the Simes test. It also shows the power of the codon models in predicting gene-wise selection for each case of the positive selection scenarios.

Table 2 indicates that the SLL test correctly predicted all datasets simulated in Case 1 as positively selected. The same predictions were made for both LRT-based tests: the M8 versus M7 and M8 versus M8A models. When the branch lengths were increased or decreased by 10-fold (Case 2 and Case 3), both being two extreme forms of sequence divergence (highly divergent or highly similar) and less likely in real data, the power of the SLL test was reduced to <80 % under $\theta_c = 0.075$ for the nominal $\alpha = 0.05$. However, if we set $\theta_c = 0.05$ to reflect the low FP in the site-wise predictions, then the powers for the two cases were increased to 97 and 96 %, respectively, which were slightly better than the LRT tests. For the very weak selection case (Case 4) all tests had much reduced power and under the more stringent criteria ($\theta_c = 0.075$) the SLL test had better power (27 %) than the M8 + M8A test (only 10 %). For the less weak selection case (Case 5) the SLL had a slightly lower power (75 %) under the stringent criteria than the M8/M8A test (80 %). However, using a less-stringent criterion ($\theta_c = 0.05$) the SLL test had much higher power, being 80 % for Case 4 and 98 % for Case 5, higher than the more powerful M8 + M7 test. For Case 6 both LRT tests showed no predictive power, consistent with the fact that the simulated data violated the assumption for the codon models that rates of synonymous substitutions be the same across the codon sites. The SLL test, however, retained some power for both the stringent and less-stringent θ_c criteria and it reached a power of 60 % for the latter.

Table 1 Number of sites falsely identified as under positive selection

Method	Cutoff	Wrong sites	%	Range (%)
SLL	0.95	1661	3	1–18 (0.3–6)
	0.99	313	0.5	0–6 (0–2)
M8 (NEB)	0.95	24402	41	0–300 (0–100)
	0.99	21750	36	0–300 (0–100)
M8 (BEB)	0.95	5	0.01	0–2 (0–0.66)
	0.99	0	0	0–0 (0–0)

Two hundred datasets of 300 codon sites were simulated under neutral evolution ($\omega = 1$) and analyzed

Power of Site-Wise Prediction

To compare the power to detect positive selection at individual sites, we plotted ROC curves averaged over the 100 replicated datasets for each case. For comparative purposes, we plotted the power of SLL and two codon models (M3 and M8), which estimate empirical Bayes posterior probabilities that a site belongs to a positive selection class. Figure 2a shows the ROC curves for the four predictors (SLL, NEB-M3, NEB-M8, and BEB-M8) in **Case 1**. All four predictors showed good power in predicting the positively selected sites. For example, at a FP rate of 0.05, the TP rates for the four predictors were 0.88 (SLL), 0.80 (NEB/M3), 0.95 (NEB/M8), and 0.95 (BEB/M8); at a FP rate of 0.20, the TP rates were increased to 0.97, 0.97, 0.99, and 0.99, respectively. Overall, Fig. 2a shows the BEB and NEB predictors under M8 had higher power than the SLL predictor, which in turn had higher power than the NEB predictor under M3, especially at the more useful low FP range (FP < 0.10).

Case 2 is a case where the simulated sequences were highly divergent and nucleotide substitutions were saturated. The average ROC curves for the 100 replicated datasets are shown in Fig. 2b. The BEB predictor under M8 obtained the highest power, followed by the NEB-M8 and the SLL, while the NEB predictor under M3 again showed the much lower power than the other models. Since the generating model is M3, it may seem surprising that

Table 2 Power of the SLL tests and the codon models

Positive selection case	Number of positive sets predicted under SLL			Number of positive sets predicted under LRT (M8 + M7/M8 + M8A)[b]
	Binomial test[a]	Simes test	Combined and unique sets[a]	
1) Original conditions from Lysin data	100 (100)	73	100 (100)	100/100
2) Branch lengths increased 10 fold	72 (97)	13	73 (97)	96/91
3) Branch lengths decreased 10 fold	67 (96)	31	79 (96)	92/95
4) Weak conditions 1 ($\omega 3 = 1.5$, $p3 = 0.05$)	16 (76)	13	27 (80)	39/10
5) Weak conditions 2 ($\omega 3 = 1.5$, $p3 = 0.269$)	71 (98)	16	75 (98)	95/80
6) Concatenated 3 M0 datasets	10 (57)	5	15 (60)	2/4

For each case, 100 datasets each of 200 codon sites were simulated and analyzed

[a] The *first number* was based on $\theta_c = 0.075$ corresponding to the standard site-wise test $\alpha = 0.05$ and the *second number* in *brackets* was based on $\theta_c = 0.05$ corresponding to a site-wise test $\alpha = 0.03$ (see *text* for details)

[b] The M8 + M7 test has two degrees of freedom ($\chi^2_{2,\,0.05} = 5.991$) and the M8 + M8A test has one degree of freedom ($\chi^2_{1\,,0.05} = 3.84$)

posterior probabilities from M8 did better than M3. The reason seems to be better estimation of the ω for the positive selection class under M8. The true values used for simulating the positive selection class were $\omega = 3.065$ and $p = 0.269$. The M3 model had a mean square error (MSE) of ω estimation being 0.21 for 50 of the data sets but the MSE was 3.22 for the other 50, whereas the M8 model had a MSE for ω to be 0.14 for 89 of the data sets and 3.19 for the other 11. The incorrectly estimated parameters under both M3 and M8 showed the same pattern: the ω for the positive selection class was underestimated (the estimated ω was <1.46 for these datasets) and its corresponding weight was overestimated ($p \geq 0.58$). Because the M3 model mis-estimated the parameters in half of the datasets the average TP rates for the M3 NEB predictor were lowest among the four predictors in Fig. 2b. The SLL predictor, not based on the prediction of the ω and p parameters, showed a fairly good ROC curve. These results indicate that the SLL predictor and the M8 predictors are good at predicting positive selection sites even when the sequence data are highly saturated with multiple substitutions.

Case 3 is a case where the simulated sequences were very similar, which proved to be difficult for the codon models (Anisimova et al. 2002; Suzuki and Nei 2004). The power of all methods at various FP rate points was reduced (Fig. 2c) compared to the data simulated under the original tree (Fig. 2a), underscoring the difficulty in predicting positively selected sites for all methods when the sequences have little variation. The M8 BEB predictor was slightly better than SLL, which was in turn slightly better than the M8 NEB predictor and again the M3 NEB was worst.

Conditions of very weak selection were evaluated in **Case 4**. Previous studies have shown that the power of the codon models in predicting positively selected sites is reduced when data are simulated under weak selection conditions (Anisimova et al. 2002; Massingham and Goldman 2005). Figure 2d shows the average ROC curves for the four predictors in this case. All methods showed much lower power compared with that in Case 1. The M8 BEB predictor performed best and the M3 NEB predictor did worst; the M8 NEB was slightly better than the SLL in some range of the FP rates. **Case 5** has the same ω value (1.5) in the positive selection class but its proportion (0.269) is much higher than in Case 4. The average ROC curves (Fig. 2e) shows the difference between the models are small although the two M8 predictors (BEB and NEB were slightly better than the M3 NEB and SLL.

In **Case 6** the data were concatenated from three datasets each simulated separately under the M0 model with $\omega = 0.08$, 0.91 and 3.06, respectively. The combined data violate the standard assumption of codon site models that the synonymous substitution rates are constant and only the non-synonymous substitution rates vary across the different ω classes (Smith and Hurst 1999; Pond and Muse 2005). Under these conditions the codon models are misspecified and may not perform well. Indeed, the codon LRT tests detected very few datasets under positive selection (Table 2). On predicting positively selected sites, the M3 and M8 models only identified six and four datasets, respectively, that have a class with $\omega > 1$ having weight >0.05. Figure 2f shows the four ROC curves for the SLL, M8 NEB, M8 BEB, and M3 NEB predictors. The ROC curve for the SLL predictor was averaged over all 100 replicated datasets, while the curves for the M8 and M3 predictors were averaged over the four and six datasets that had a significant proportion of sites (weight > 0.05) in the class of $\omega > 1$. While the SLL predictor reached a power of 0.55 at a FP rate of 0.2 in detecting the positive sites the codon models had no power even in the few datasets that they had predicted to have a $\omega > 1$ class.

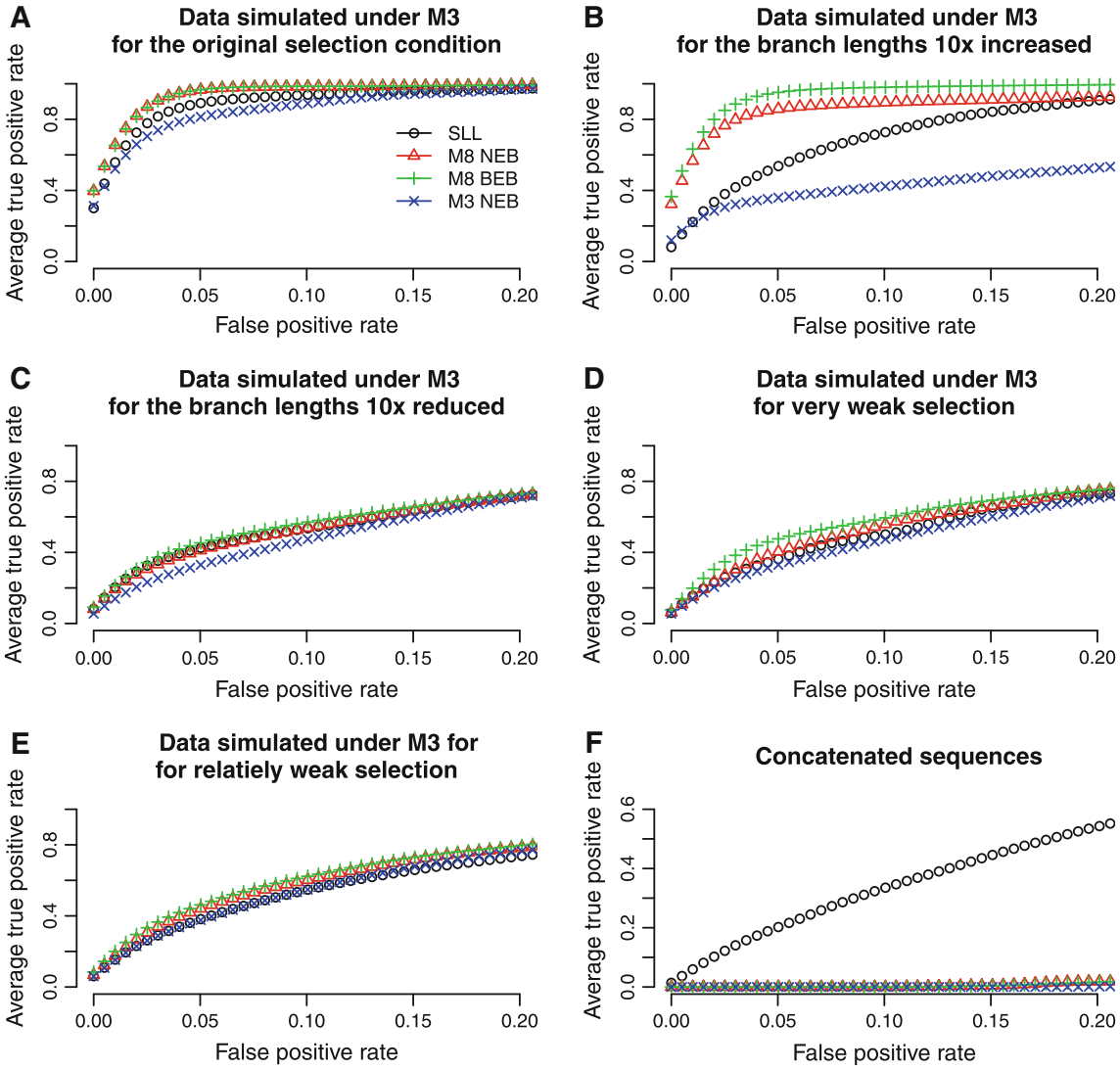

Fig. 2 **a–f** are the average ROC curves showing the power of several predictors for positively selected sites on the 100 replicated datasets simulated under each of the six cases of positive selection. In (**f**) the ROC curves for the M3 and M8 models were averaged over the six and four datasets that were predicted to have significant amount of positively selected sites under the two models (their proportions were >0.05). The predictors (see **a** for legends) include SLL; the naïve empirical Bayes probabilities estimated under M3 (M3 NEB) and under M8 (M8 NEB); and the Bayes empirical Bayes probabilities estimated under M8 (M8 BEB)

Empirical Case Studies

The ten real datasets listed in Table 3 have relatively minor codon usage biases as measured by their effective number of codons (Wright 1990) that range from 38.2 (vertebrate β–globin and HIV-1 *vif* genes) to 53.6 (Japanese encephalitis envelop gene). We applied the SLL test to the ten datasets and for comparison we also used the codon models (M7, M8, and M8A) to the same data. Table 3 lists the numbers of taxa and sites in each dataset; the estimated θ and its critical value θ_c at the $\alpha = 0.05$ level for the binomial test; the Simes test; the positively selected sites for the positive genes predicted under the SLL; the LRT tests and the positively selected sites with the BEB and NEB posterior probabilities >95 % estimated under M8. Since the SLL-binomial test criteria were rather stringent as seen in the simulation studies for both type I error rates and power analyses, the results for the two datasets (HIV-1 *pol* and HIV-1 *env*-V3) were marked as ± as the estimated θ were only slightly smaller than θ_c and indeed the Simes test predicted HIV-1 *pol* to be positive. The HIV-1 *pol* and vertebrate β–globin data were previously analyzed with a site-wise likelihood ratio test (Massingham and Goldman 2005). The HIV-1 *env*-V3 data were also pre-analyzed with several other methods for detecting positive selection including a random effects likelihood method which

Table 3 The SLL and codon LRT tests applied to ten real datasets

Dataset (Number of taxa/Number of sites)	SLL			LRT	
	Gene level selection		Positive sites[a]	Gene level selection	Positive sites[b]
	Binomial test (θ/θ_c)	Simes test		M8 + M7/M8 + M8A	
Abalone sperm lysin (25/135)	0.09/0.08: +	−	4 9 32 33 **40** 41 44 70 74 83 86 120	+/+	BEB: 4 7 9 10 12 14 32 33 36 41 44 64 67 70 74 83 86 87 113 120 126 127 132 134 NEB: same as BEB without 127
Drosophila alcohol dehydrogenase (23/254)	0.03/0.07: −	−	(49 68 69 98 163 170 201)	−/−	None
Flavivirus E-glycoprotein (22/496)	0.03/0.07: −	−	(37 46 81 124 132 171 203 227 272 303 322 357 383 391 492)	−/−	None
Flavivirus NS-5 (18/342)	0.04/0.07: −	−	(42 61 64 66 103 106 173 176 188 189 216 238)	−/−	None
Vertebrate β–globin (17/144)	0.04/0.08: −	−	(11 50 67 85 123)	+/+	BEB: 7 123 NEB: 7 50 67 85 123 **SLR**[c]: none
HIV-1 pol (23/947)	0.05/0.06: ±	+	2 3 4 14 39 41 67 **97 223** 224 **237 264 302** 313 347 **374** 379 388 395 **399** 431 **450** 459 462 478 479 **481 488** 492 **503 535 552** 568 570 **583 650** 654 **670** 671 732 761 771 779 **782 784 816** 890 892 **894 925**	+/+	BEB: 41 67 347 379 431 459 478 479 492 671 771 779 NEB: 2 3 39 41 67 224 313 347 379 395 431 459 478 479 492 568 570 654 671 761 771 779 892 **SLR**[c]: 2 3 4 14 41 67 313 347 379 388 431 459 462 478 568 570 654 732 761 779 890
HIV-1 vif (29/192)	0.10/0.08: +	+	**22** 31 33 **37** 39 **47 48** 63 92 101 122 127 **128 132 151 154 155 159** 167	+/+	BEB: 31 33 39 63 92 101 109 122 127 167 NEB: same as BEB
HIV-1 *env*-V3 (13/91)	0.08/0.09: ±	−	26 28 **40** 51 66 **69 76**	+/+	BEB: 26 28 51 66 87 NEB: 28 66 87 **REL**[c]: 26 28 40 51 66
Japanese encephalitis env (23/500)	0.05/0.07: −	−	(33 35 36 76 126 129 138 146 153 161 166 222 227 242 253 323 327 336 366 387 399 434 466 486 490)	−/−	None
Human T cell lymphotropic virus *Tax* (20/181)	0.04/0.08: −	−	(2 39 53 115 146 154 166)	+/+	BEB: 2 4 39 43 53 60 62 69 81 85 92 101 108 115 146 152 154 157 161 166 181 NEB: all sites positive with $p = 1$ **Pars**[c]: none

[a] The positive sites were predicted under the SLL test; those also predicted under the Benjamini–Hochberg procedure for multiple tests corrections (see Methods) were underscored. The sites whose BEB or NEB posterior probabilities were <95 % are in *bold font*. The other sites had BEB or NEB probabilities >95 %. Sites in *brackets* are those that had site lnL smaller than the 5th percentile of the simulated lnL distribution under no selection but the whole data did not pass the SLL test

[b] Sites with BEB and NEB > 95 % are shown

[c] Results obtained from previous analyses of these data sets

SLR site-wise likelihood ratio test (Massingham and Goldman 2005), *REL* random effects likelihood (Pond and Frost 2005), *Pars* parsimony-based analysis (Suzuki and Nei 2004)

considered both non-synonymous and synonymous rate variations across sites (Pond and Frost 2005). The *Tax* data were analyzed with a parsimony-based analysis (Suzuki and Nei 2004). The results from the previous methods for these four datasets were also included in Table 3.

The SLL method made the same predictions as the codon models in eight of the ten datasets at the gene level. They made different predictions in the other two datasets (the vertebrate β–globin and Human T cell lymphotropic virus *Tax* genes), where positive selection was inferred under the codon models but not SLL. As will be discussed in the following section, this was not due to a weaker power in SLL and, in fact, at least for the *Tax* gene data, the SLL prediction appeared more reasonable. The predictions of positively selected sites were largely similar for both methods, although there were some positive sites (e.g., site 40 and site 87 of the HIV-1 *envelop* gene V3 region) that were predicted under SLL but not under the codon models or vice versa. This will also be discussed below.

Discussion

We have shown that there is a negative correlation between the site log-likelihood in an amino acid alignment and the probability that a site has experienced positive selection.

There are several possible explanations of why this correlation exists. First, it is expected that sites undergoing positive selection (especially those of diversifying selection associated with host-parasite co-evolution and sexual conflict at the molecular level, for example) will necessarily be more variable on average (Hayes et al. 2010). This can be demonstrated for example using a simple three taxon star tree and the WAG substitution model (see Supplementary Figure S1). As shown in Figure S1, site patterns with more states tend to have lower site-likelihoods than site patterns with fewer states. Indeed, there is a strong negative linear correlation between the log-likelihoods and parsimony scores across the amino acid sites (regression $R^2 = 0.97$); yet the correlation is much weaker ($R^2 = 0.41$) between the site-wise log-likelihoods and the consistency indices (Kluge and Farris 1969) indicating low likelihoods are more related to site variability than to homoplasy per se. Another factor responsible for this correlation is related to the "radical" nature of substitutions that occur on the amino acid level at positively selected sites. Although most codon models such as those used here do not directly account for "radical" versus "conservative" changes on the amino acid level, the number of codon changes required for more radical amino acid substitutions tends to be greater. For example, if one plots the exchangeabilities between amino acid types in the WAG matrix (Whelan and Goldman 2001) versus the minimum number of codon substitutions required for a given substitution in a codon model, it is clear that amino acid interchanges with low exchangeabilities (i.e., more "radical" changes) tend to require more substitutions on the nucleotide level (Figure S2). As a result it is not surprising that sites where a large number of non-synonymous substitutions occurred have an elevated probability of being in a positively selected site class and these sites, by virtue of displaying more radical amino acid interchanges, will tend to have lower site log-likelihoods. Furthermore, some studies have pointed out positively selected sites are more likely to appear in the solvent exposed areas of protein surfaces, which also tend to be more variable (e.g., Osorio et al. 2007; Meyer and Wilke 2012).

Regardless of the reasons for the correlations, we have demonstrated with simulations that site-wise log-likelihood tests (SLL) can be used to predict genes under positive selection and detect the positively selected sites. For the 200 datasets simulated under strict neutral evolution (the hardest case to distinguish from positive selection), the SLL-binomial test had a very low FP (0 %) and the Simes test had a FP of 6.5 % at the gene level. Both appeared better than the M8 + M8A LRT test, which had high FP of 19.5 and 8.5 % under a ½ χ_0^2 + ½ χ_1^2 mixture and a χ_1^2 criterion, respectively. Another commonly used LRT test, comparing M8 and M7 models, had a FP of 2.5 %. For the same data, the FP for predicting sites under positive selection was 3 % for the SLL test and it was very high (41 %) under the NEB posterior probabilities and very small under the BEB probabilities from the M8 model. A number of previous studies have shown that, the NEB prediction under the M8 model had a very high FP (35 %) if the datasets were not prescreened for being under positive selection at the gene level (e.g., Anisimova et al. 2002; Massingham and Goldman 2005). Given the continuous debate in the literature about the FP of codon-based models (Suzuki and Nei 2004; Hughes 2007; Zhai et al. 2012) the low FP in the SLL tests are appreciable.

We further studied the power of the SLL tests in six cases of positive selection simulated based on a lysin tree. For the data simulated under the same conditions (including the ω parameters and branch lengths) as estimated from the real lysin data, the SLL test showed perfect power in predicting the datasets under positive selection (Table 2). The SLL test also had high power in predicting the positively selected sites as shown in the ROC curve plotting the TP to the FP, similar to the site-wise Bayes predictions under the M8 model (Fig. 2a). For the simulated cases where the sequences were very similar or very divergent or had very weak signal of positive selection the SLL tests had less power in predicting datasets under positive selection than seen in the LRT tests (Table 2). However, the ROC curves showing the power in predicting positively selected sites were not much different from the NEB predictor under M8 and often better than the M3-based NEB predictor (Fig. 2b–e). This suggested the SLL-binomial test based on the θ_c calculated under the nominal $\alpha = 0.05$ (Eq. 1) was too stringent for deciding positive selection at the gene level. Setting $\alpha = 0.03$ in Eq. 1, which was the FP of the SLL in the site-wise prediction (see Table 1), the proportion of the datasets predicted to be under positive selection was much larger in each of the four cases and better than the power of the commonly used M8 + M7 test (Table 2). For the sixth case where the datasets were concatenated data from three sub-datasets each simulated under M0 with one of them being a positive selection class, the assumption for the codon models that synonymous substitutions rates are constant across the sites was violated. Such violations are possible with real data. For instance, a multi-gene dataset in a phylogenomic study may contain genes with synonymous substitution rates varying across sites or loci, a phenomenon that could occur, for example, when codon usage in highly expressed genes is subject to selection (Akashi and Eyre-Walker 1998) or there are strong genetic hitchhiking effects (Barton 2000). Both M8 + M7 and M8 + M8A tests detected very few sets under positive selection while the SLL identified 60 positive sets under the less stringent θ_c criterion. Furthermore, both NEB and BEB based ROC

curves under the M3 and M8 models had no power in detecting positively selected sites even for the very few datasets that were estimated to have a class of $\omega > 1$ and $p > 0.05$, whereas the SLL curve showed some power and the average TP rate reached 0.55 at a FP rate of 0.2 for all 100 datasets (Fig. 2f). This demonstrates that the SLL method can partially alleviate the problem associated with variation in synonymous substitution rates while the standard codon site model did not, although newer codon models (Rubinstein et al. 2011) that take into account this variation were not tested here.

The results of the analyses of the ten empirical datasets indicate that the SLL test usually made similar predictions to the codon LRT tests. However, the two methods made different predictions for the vertebrate β-globin and the HTLV *Tax* data sets. Both data sets were predicted to have experienced positive selection using LRT tests but not so under the SLL test. For the β-globin data the non-positive prediction by SLL may be due to its relatively lower power for gene-wise testing, although it is unknown what the correct prediction is in this case as some studies have suggested there are no positively selected sites in this gene (Massingham and Goldman 2005). The *Tax* gene data, however, clearly showed the benefit that the SLL can make more reasonable gene-wise prediction in some case. The *Tax* data was originally used to argue that the codon site models can sometimes generate false predictions of positively selected sites with high confidence even though the proteins under consideration were possibly not subject to positive selection (Suzuki and Nei 2004). The data were comprised of 20 taxa and 181 codon sites, 87 % of which were constant across all sequences. Of the remainder 23 sites, 2 had a single sequence with a synonymous change and 21 had a single sequence with a non-synonymous change, 19 of which were nucleotide transversions and two were transitions. Both SLL and a parsimony-based method (Suzuki and Nei 2004) predict that the gene has not experienced positive selection. However, all codon models (M0, M3 and M8) predicted very strong positive selection in this gene ($\omega = 4.87$ under M0) and all 181 sites were estimated by the NEB predictions to be under positive selection with 100 % posterior probability (Table 3). The BEB under M8, although less liberal, predicted all 21 sites with a single non-synonymous change to be positively selected with high confidence.

To determine thresholds for the SLL test, a null distribution is obtained from a large number of codons (20,000 in all analyses in the paper) simulated under neutral evolution conditions using the M8A model, which takes the same parameter values from the original data estimated under a M8A model with equal codon frequencies (all 1/61, the F0 model of codon frequencies in PAML). The simulated codon sequences are then analyzed under a standard amino acid model with a discrete Gamma distribution (WAG + F + G in this study) and the 5th percentile of the site log-likelihood distribution is used as a threshold to compare with the site log-likelihoods from the original data estimated under the same amino acid model. To allow unequal frequencies of codons one can determine thresholds as above with M8A but with codon frequencies modeled using nucleotide frequencies at the three codon positions (i.e., the F3 × 4 model in PAML). We found, however, that the type I error of the SLL tests were fairly robust to the assumption of equal codon frequencies. To test these we reconsidered simulations for the type 1 error rates described in the earlier sections (see Table 1). We simulated the data under M0 with the codon frequencies being from the F3 × 4 model applied to the lysin data, rather than using equal codon frequencies as in the original simulations for testing type 1 errors. The resulting SLL tests using M8A under F0 codon frequencies to determine the thresholds had a type 1 error rate of 0.051 at the site-wise level and it was 0.049 at the gene-wise level (combined binomial and Simes tests). In addition, we found that the performance of the gene-wise test was degraded when F3 × 4 frequencies were used to determine thresholds (Table S1 in Supplemental Material 3). Other choices of null simulating models are also possible and should be investigated in the future. They include simulating a large number of codon sites under the M0 model with $\omega = 1$ or under an unrestricted empirical codon model (Kosiol et al. 2007; Doron-Faigenboim and Pupko 2007) and determining a corresponding amino acid site log-likelihood threshold under an amino acid model. The latter approach may be of particular interest as this way we can have a fair comparison between the SLL and the codon models as data generated under an empirical codon model will not bias the result favoring the mechanistic codon models which were used to simulate all the data in the current analyses.

Although we have demonstrated the usefulness of the SLL approach for detecting positive selection from a sequence alignment through simulation and several case examples, it is worthwhile to consider the impact of model misspecification on the effectiveness of this approach. On the one hand, model misspecification may improve the identification of true positively selected sites, as the positively selected sites are more variable and often contain radical amino acid changes that are not expected from the standard amino acid model and so will have lower probability of occurring. On the other hand, many forms of model misspecification at a site coupled with relatively high site-rates are likely to manifest in very low site likelihoods in amino acid models which could mislead one into inferring that a site is under positive selection, even if it was not. In codon models the synonymous changes serve as a kind of "barometer" so that, in theory, positive selection

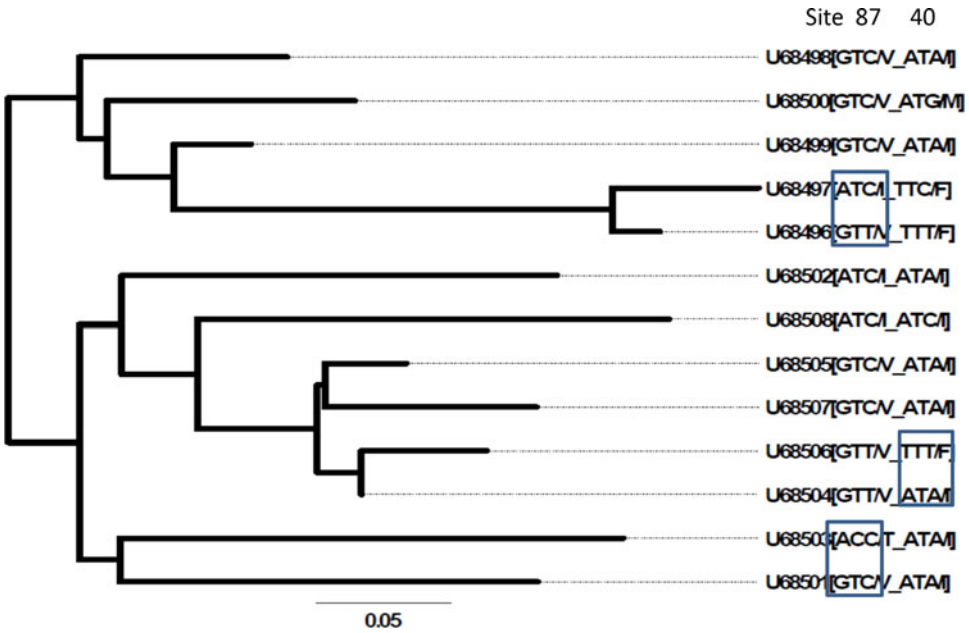

Fig. 3 The phylogenetic tree of HIV-1 *env* gene V3 region estimated under WAGF + Γ, with codons and amino acids at site 87 and site 40 indicated in the *brackets* after the taxon names. The first codon and the following one-letter amino acid residue, *separated by /*, are for site 87; the second codon and its one-letter amino acid residue (*also separated by /*) are for site 40. The codons between two neighboring taxa containing changes in 2 nucleotides are *boxed*

is inferred only when the rate of amino acid changes is occurring faster than the rate of synonymous changes in codons. However, if some non-synonymous sites are changing in a way that does not obey any model of the best-estimated codon rate matrices, then it is possible that this will lead to erroneously better fit to the class with large dN/dS ratios where rare unexpected nonsynonymous changes happen more often. An example of the latter could be an excess of changes between amino acids that require more than one non-synonymous nucleotide change in the codon. For mechanistic codon models, these are supposed to have a zero instantaneous rate and thus can only be explained by two or more sequential changes, but it is possible that in nature some codons in some genes are prone to "double" mutations that allow this to happen in one step (Kosiol et al. 2007; Doron-Faigenboim and Pupko 2007).

To illustrate this point, we consider the sequence pattern of two codon sites (site 87 and site 40) in the HIV-1 *envelope* gene V3 region data. Site 87 was predicted to be a positively selected site with high posterior probability support (>99 %) under the M8 model, but it was not picked as such a site under the SLL. Site 40, however, was detected as a strong positively selected site under the SLL but was not so estimated under the codon models (the posterior probability was 0.73 under M8). Site 87 consisted of the amino acids V, I, and a single T, whereas site 40 consisted of I, F, and a single M. The rate in the WAG + F matrix from V to I is 0.52 which is comparable to the rate from M to I (0.5) but the rate from I to V is 0.95 which is much larger than the rate from I to M (0.01). The reciprocal rates between I and T and between V and T, which are between 0.1 and 0.18, are also greater than the rates between F and I and between F and M, which are between 0.01and 0.13. Consequently, it is not surprising that site 40 has smaller amino acid site likelihood than site 87. However, at site 87 there are two pairs of sister taxa that have codons differing by two nucleotides, while site 40 has only one pair of sister taxa with codons differing by two nucleotides (Fig. 3), which explains why the codon model predicted site 87 as a strong positively selected site but not site 40. It is intriguing to note that Pond and Frost (2005) whose methods account for rate variations both among synonymous and non-synonymous codon sites, also found site 40 to have a relatively high posterior probability of positive selection ($p = 0.899$) but did not for site 87. Moreover, site 40 is within the amino acid stretch (sites 37–51) of the envelope protein that interacts with monoclonal antibodies (Yamaguchi and Gojobori 1997) and the whole consensus V3 loop (sites 27 through 60) enclosed by the disulfide bridge linking a pair of cysteine residues also contains site 28 and site 51 that are predicted to be positively selected by both SLL and codon models.

Conclusions

We used both simulations and empirical case studies to show the amino acid site likelihood score estimated under the standard amino acid models can be used to predict proteins under positive selection with a low FP and a power comparable to the standard codon models. For data that fit the codon site models well, the amino acid SLL approach predicts sites under positive selection nearly as well as

using the codon site models to predict the Bayes posterior probabilities of sites being in a positive selection class. We also gave an example where the codon models generate many false-positive predictions with high confidence for a very conserved dataset with few and peculiar substitution patterns. In this case the SLL test made a more reasonable prediction of no positive selection in the data. While our method should not be considered as superior to the widely used codon models, it provides an alternative way to detect positive selection that is robust to some forms of codon model misspecification as shown in Case 6 of the simulation studies. In circumstances where the assumptions of the codon models are violated (e.g., the rate of synonymous substitutions is not constant across sites or the codon usage is changing drastically over the tree) or these models cannot be used (e.g., where the genetic code differs across sequences in the data), the SLL method can still be utilized and may provide a more robust result.

Acknowledgments We thank Tal Pupko and two anonymous reviewers for insightful comments. This study was supported by Discovery grants awarded to AJR and ES by the Natural Sciences and Engineering Research Council of Canada (NSERC). AJR acknowledges support from the Canadian Institute for Advanced Research Program in Integrated Microbial Diversity and the Canada Research Chairs program. HCW is currently supported by a CGEB postdoctoral fellowship from the Tula Foundation.

References

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

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

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

Barton NH (2000) Genetic hitchhiking. Phil Trans Royal Soc Lond B Biol Sci 355:1553–1562

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 57(1):289–300

Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24:388–397

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

Green DM, Swets JM (1966) Signal detection theory and psychophysics. Wiley, New York

Hanada K, Shiu SH, Li WH (2007) The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. Mol Biol Evol 24:2235–2241

Hayes ML, Eytan RI, Hellberg ME (2010) High amino acid diversity and positive selection at a putative coral immunity gene (*tachylectin*-2). BMC Evol Biol 10:150

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364–373

Hughes AL, Hughes MK (1993) Adaptive evolution in the rat olfactory receptor gene family. J Mol Evol 36:249–254

Hurst LD, Pal C (2001) Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet 17:62–65

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Jordan G, Goldman N (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29:1125–1139

Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of Anurans. Syst Zool 18:1–32

Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58

Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24:1464–1479

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

Meyer AG, Wilke CO (2012) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44

Nickel GC, Tefft DL, Goglin K, Adams MD (2008) An empirical test for branch-specific positive selection. Genetics 179:2183–2193

Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA 106:6700–6705

Osorio DS, Antunes A, Ramos MJ (2007) Structural and functional implications of positive selection at the primate angiogenin gene. BMC Evol Biol 7:167

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

Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22:2375–2385

Privman E, Penn O, Pupko T (2012) Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol 29:1–5

Pupko T, Sharan R, Hasegawa M, Shamir R, Graur D (2003) Detecting excess radical replacements in phylogenetic trees. Gene 13:127–135

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna http://www.R-project.org

Rubinstein ND, Mayrose I, Doron-Faigenboim A, Pupko T (2011) Evolutionary models accounting for layers of selection in protein coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297–3308

Shapiro BJ, Alm E (2009) The slow:fast substitution ratio reveals changing patterns of natural selection in γ-proteobacterial genomes. ISME J 3:1180–1192

Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Smith NGC, Hurst LD (1999) The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152:661–673

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

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

Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20:18–20

Tang H, Wu C-I (2006) A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. Mol Biol Evol 23:372–379

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Wong WSW, Yang A, 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

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29

Yamaguchi Y, Gojobori T (1997) Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc Natl Acad Sci USA 94:1264–1269

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford

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

Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228

Yang Z, Nielsen R, Goldman N, Pedersen AM (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 WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zhai W, Nielsen R, Goldman N, Yang Z (2012) Looking for Darwin in genomic sequences—validity and success of statistical methods. Mol Biol Evol 29:2889–2893

Zhang J (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56–68

Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27:1912–1922