Supplementary Information

Not so different after all: a comparison of methods for detecting amino-acid sites under selection

Sergei L. Kosakovsky Pond and Simon D.W. Frost*

Antiviral Research Center, University of California San Diego, San Diego, California, 92103

Research Article

Running title: Methods for Detection Of Selection

Keywords: positive and negative selection, codon substitution models, substitution rates, parallel algorithms

* Corresponding author

Antiviral Research Center

University of California San Diego

150 W Washington Street, Suite 100

San Diego, California, 92103

E-mail: sdfrost@ucsd.edu

Phone: +1 619 543 8080

Fax: +1 619 298 0177

Extended Binomial Distribution

Building upon the notion of the *almost binomial* distribution supported on $0 \ldots [N]+1$ ($[N]$ denotes the largest integer not greater than N) and defined by

$$Pr_{ab}\{X = k\} = \begin{cases} \begin{pmatrix} N \\ k \end{pmatrix} P^k (1-P)^{N-k}, & k = 0 \ldots [N], \\ 1 - \sum_{l=1}^{[N]} \begin{pmatrix} N \\ l \end{pmatrix} P^l (1-P)^{N-l}, & k = [N]+1, \end{cases}$$

we let $EBin(N, P)$ have piecewise constant density equal to $Pr_{ab}\{X = k+1\}$ for $x \in [k, k+1)$ and to place the point mass of $Pr_{ab}\{X = 0\}$ at 0. Thus, $EBin(N, P)$ is supported on the interval $[0, [N]+1]$.

If N is an integer, $Y \sim EBin(N, P)$ and $X \sim Bin(N, P)$, then it is easy to see that for any integers $0 \leq k < m \leq N$, $Pr\{X \in [k, m]\} = Pr\{Y \in [k, m]\}$, and, therefore the tails of the distributions are equal for integer values k, N.

One can assess whether or not the extended binomial distribution is a good approximation to the null distribution of substitution counts on real data by simulating the null distribution parametrically. Having fitted a codon model, and given any ancestral codon at the root of the tree, one can generate the distribution of synonymous and non-synonymous counts parametrically, assuming neutrality, i.e. $dN = dS$, holding other model parameters (nucleotide rate biases and branch lengths) at their MLE values. As a byproduct of maximum likelihood codon-based ancestral state reconstruction, it is easy to obtain the relative support of each possible codon at the root for every codon site in the alignment. It is then straightforward to tabulate the expected number of synonymous substitutions S at a site, conditioned on the total number of inferred substitutions N. We simulate a large number of alignment columns, e.g. 50000 per root codon state. This only has to be done once for entire alignment, as the only component which must be recomputed for each site is the re-weighting based on sup-

port probabilities for each root state. We then consider only those simulations that contain exactly N total codon substitutions. The p value for positive selection at a site can then be determined by looking up the simulated probability that $S < S_s$, where S_s is the number of synonymous substitutions inferred by a counting method. Such an approach is much more computationally intensive than SLAC, and it is comparable to FEL in terms of computer time required.

On HIV-1 alignments analyzed in this paper, the EBin approximation appears to work reasonably well. While there are noticeable differences in p-values overall (see Figure 6, for example), the sites inferred to be under positive selection by both approaches for all three data sets nearly coincide, when SLAC is used. Clearly, the EBin is simply a useful approximation, and there may be practical scenarios under which it is not appropriate, but our limited simulations and the broad agreement between counting and likelihood methods suggest that the approximation is useful. Further studies of the conditions under which the approximation is warranted would clearly be beneficial.

Effect of parameter estimation errors on REL

To assess estimation errors in parameter estimates, we calculated the confidence intervals for maximum likelihood parameter estimates using profile likelihood. These errors are likely to be underestimates of the true error, as nuisance parameters are fixed at their maximum likelihood values when calculating profile likelihoods. We investigated the robustness of our results to changing the estimates of rate parameters. For our *env* dataset, maximum likelihood estimates of the rate distributions for synonymous rates α were $\alpha_1 = 0.0$ ($P = 0.55$), $\alpha_2 = 2.2$ ($P = 0.45$) and for non-synonymous rates β: $\beta_1 = 0.0$ ($P = 0.29$), $\beta_2 = 0.75$ ($P = 0.54$), $\beta_3 = 3.5$ ($P = 0.17$). Using more conservative values from 95% profile likelihood confidence intervals for the above distributions, we observed that changing the β distribution to $\beta_1 = 0.0$ ($P = 0.4$), $\beta_2 = 1.71$ ($P = 0.48$), $\beta_3 = 4.78$ ($P = 0.12$) lowered the log-likelihood

by 2.4, which places altered parameter estimates inside a 95% confidence interval around the MLE. However, using this modified distribution of β to perform the empirical Bayes analysis yielded lower Bayes factors for putatively positively selected sites: 10 for codon 26, 387 for codon 28, 26 for codon 51, 40 for codon 66 and 8 for codon 66 (compare with Table 2). Thus, only a single site (codon 28) remained above the threshold of 50. Note that codon 28 is the only one classified as positively selected by the other methods.

Effect of using an incorrect model of nucleotide substitution

This effect can be noticeable for larger data sets - for instance, when we ran SLAC with HKY85 in place of (012232) model (notation of Muse (1999)) on 297 HIV-1 drug naïve pol sequences discussed previously, we found that the log-likelihood score was significantly worse for $MG94 \times HKY85$ (-18423.3), than for $MG94 \times (012232)$ (-18333.4). Also, $MG94 \times HKY85$ "lost" one of the sites found to be under positive selection by $MG94 \times (012232)$: codon 102, and also yielded larger p-values (likely due to weaker power) for all but one of the remaining sites found under selection (Table 4).

Proof of WAC equivalence claim

We wish to compute the expectation of a function $F(A)$ which depends on the assignment of ancestral states A to internal nodes of tree T, whose leaves are labeled with observed characters D. The probability distribution on the set of all possible ancestral states $\mathcal{A}$ is induced by the standard phylogenetic likelihood function,

$$Pr\{A\} = \frac{L(D|A,\theta)}{\sum_{a\in\mathcal{A}} L(D|a,\theta)},$$

where θ is introduced to denote the (fixed) values of all other model parameters.

Furthermore, assume that F is additive along branches, i.e. $F(A) = \sum_b F(b_p, b_d)$, where b indexes all tree branches, while b_p and b_d denote characters at the parent and child end of the branch, respectively. Note that all the quantities computed by our counting algorithms possess the branch-additive property.

For clarity, but without loss of generality, we explain the equivalence on an example tree shown in Figure 5. An ancestral state for the example tree is comprised of four internal node labels: c_6, c_7, c_8, c_9. We use the notation $Q^k_{c,d}(t;\theta)$ to denote the probability of substituting character c with character d along branch k over time t given model parameters θ. $\mathcal{C}$ denotes the set of all possible character states, in our case - codons which are not stop codons.

By definition,

$$\begin{aligned}
E\left[F(A)\right] &= \sum_{c_9\in\mathcal{C}}\sum_{c_8\in\mathcal{C}}\sum_{c_7\in\mathcal{C}}\sum_{c_6\in\mathcal{C}}\left[F(c_9,A)+F(c_9,c_8)+F(c_8,C)+\right.\\
&\quad \left.F(c_8,c_7)+F(c_7,c_6)+F(c_7,T)+F(c_6,T)+F(c_6,T)\right]Pr\{A\}\\
&= \frac{1}{\sum_{a\in\mathcal{A}} L\left(D|a,\theta\right)}\sum_{c_9\in\mathcal{C}}\pi(c_9)Q^1_{c_9,A}(t_1;\theta)\sum_{c_8\in\mathcal{C}}Q^8_{c_9,c_8}(t_8;\theta)Q^2_{c_8,C}(t_2;\theta)\times\\
&\quad \sum_{c_7\in\mathcal{C}}Q^7_{c_8,c_7}(t_7;\theta)Q^3_{c_7,T}(t_3;\theta)\sum_{c_6\in\mathcal{C}}Q^6_{c_7,c_6}(t_6;\theta)Q^4_{c_6,T}(t_4;\theta)Q^5_{c_6,T}(t_5;\theta)\times\\
&\quad [F(c_9,A)+F(c_9,c_8)+F(c_8,C)+F(c_8,c_7)+F(c_7,c_6)+\\
&\quad F(c_7,T)+F(c_6,T)+F(c_6,T)]
\end{aligned}$$

If one distributes the innermost sum, one is left with a combinatorial size sum for each branch, which can be computed using the WAC algorithm for that branch. Consider the term which included by, for example, $F(c_8, c_7)$:

$$\sum_{c_8\in\mathcal{C}}\sum_{c_7\in\mathcal{C}}F(c_8,c_7)\left[\left(\sum_{c_9\in\mathcal{C}}\pi(c_9)Q^1_{c_9,A}(t_1;\theta)Q^8_{c_9,c_8}(t_8;\theta)Q^2_{c_8,C}(t_2;\theta)\times\right.\right.$$
$$\left.\left.Q^7_{c_8,c_7}(t_7;\theta)Q^3_{c_7,T}(t_3;\theta)\sum_{c_6\in\mathcal{C}}Q^6_{c_7,c_6}(t_6;\theta)Q^4_{c_6,T}(t_4;\theta)Q^5_{c_6,T}(t_5;\theta)\right)\Big/\sum_{a\in\mathcal{A}}L\left(D|a,\theta\right)\right]$$

Upon closer examination, the expression in square brackets is the Relative Likelihood Support (as defined in the main text) for the labeling c_8, c_7 along branch b_7. This expression can be efficiently calculated using a straightforward modification of Felsenstein's pruning algorithm. The same scheme can be applied to branches of arbitrary trees. Therefore, WAC indeed computes the expected value of branch-additive quantities over all possible ancestral labelings.

Power and Type II errors in the absence of synonymous rate variation

We wanted to compare the performance of our methods with the popular M8 model (Yang et al., 2000) commonly used for selection analyses, when the data conforms to the assumption of constant synonymous rates across sites. Specifically, each alignment contained 250 codons simulated along a symmetric tree with 32 sequences, with branch lengths sampled from a mean 0.05 exponential distribution. The following distribution of rates was used:

100 negatively selected sites: 100 codons with $\alpha_s = 1$ and $\beta_s = 0.25$ ($\omega_s = 0.25$)

75 neutral sites: 75 codons with $\alpha_s = \beta_s = 1$;

75 positively selected sites: 50 codons with $\alpha_s = 1$ and $\beta_s = 2$ ($\omega_s = 2$), 25 codons with $\alpha_s = 1$ and $\beta_s = 5$ ($\omega_s = 5$)

This scenario is rather difficult for M8 to fit, because of a large number of neutral sites and a mixture of two positively selected classes; M8 only allows for a single $\omega > 1$. In practice, one does not know *a priori* what rate distribution should be fitted to the data, and M8 has become somewhat of a *de facto* standard for detecting adaptive evolution, even though its choice of rate distribution may not be the most appropriate in all cases.

ROC plots (Figure 7) suggest that REL and M8 perform very similarly, while both FEL and SLAC suffer from overfitting of synonymous rates at each site, as may be expected. The

flattening of the ROC curve for M8 is due to the fact that almost all sites are classified as positively or negatively selected with relatively high posterior probability. M8 also strongly classifies many neutral sites as selected. Other methods may be more ambiguous due to their greater flexibility to model rate distributions - see the rate recapitulation plot in Figure 7.

As for previous simulations, FEL was best able to recapitulate the rates of evolution at every site, followed by REL, M8 and, lastly SLAC. This finding suggests the general use of FEL as means for estimating the true distribution of substitution rates, and detection of possible shrinkage effects, as evident for this scenario with M8 and, to a lesser extent, with REL.

Literature Cited

Muse, S. V. 1999. Modeling the molecular evolution of HIV sequences. chap. 4, Pp. 122–152, *in* K. A. Crandall, ed. The evolution of HIV. The Johns Hopkins University Press.

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

		Counting Methods						Likelihood Methods			
Codon		*SLAC*		*WAC*		*Sampler*		*FEL*		*REL*	
20		1.87	(0.14)	1.88	(0.14)	1.87:2.18	(0.10:0.14)	**2.33**	(0.08)	0.76	(26.36; 0.7549)
35	P	**2.75**	(0.10)	**2.77**	(0.10)	2.56:3.42	(0.06:0.14)	2.03	(0.36)	**1.80**	(181.96; 0.9551)
60		2.38	(0.11)	2.37	(0.12)	2.31:2.71	(0.08:0.12)	1.94	(0.28)	**1.80**	(140.03; 0.9424)
64	P	1.29	(0.24)	1.29	(0.24)	1.29:1.29	(0.24:0.24)	**1.53**	(0.03)	**0.76**	(499.55; 0.9832)
69	P	**3.55**	(0.01)	**3.55**	(0.01)	**3.54:3.55**	(0.01:0.01)	**3.44**	(0.01)	**4.18**	(4673.93; 0.9982)
83		2.35	(0.16)	2.35	(0.17)	1.92:2.68	(0.13:0.30)	1.37	(0.55)	**1.80**	(104.50; 0.9243)
102		0.65	(0.49)	0.65	(0.49)	0.65:0.65	(0.49:0.49)	0.77	(0.11)	**0.76**	(370.54; 0.9774)
178		1.18	(0.58)	1.18	(0.58)	1.18:1.18	(0.58:0.58)	1.38	(0.78)	**1.80**	(57.40; 0.8702)
200	P	**5.46**	(0.00)	**5.49**	(0.00)	**5.46:5.78**	(0.00:0.00)	**5.47**	(0.00)	**4.18**	(4688.95; 0.9982)
207	P	**6.07**	(0.01)	**5.55**	(0.02)	**4.87:6.06**	(0.01:0.05)	**9.80**	(0.04)	**1.80**	(59.11; 0.8735)
211	P	2.44	(0.23)	2.13	(0.27)	2.43:2.74	(0.23:0.23)	**4.50**	(0.04)	**1.80**	(115.61; 0.9311)
215	P	**3.24**	(0.02)	**3.24**	(0.02)	**3.23:3.24**	(0.02:0.02)	**3.34**	(0.00)	**4.18**	(12669.30; 0.9993)

Table 1. Positively selected sites in HIV-1 Reverse Transcriptase of AZT treated patients identified by at least one of the methods. The first number for every method is an appropriately scaled $dN - dS$, so that they are directly comparable. The number in parentheses show p-values for the appropriate test, and the Bayes factor values for the REL method; posterior probabilities are also included for reference purposes, although they are not used in site classification. The entries for the Sampler method show 95% quantiles for the distribution of $dN - dS$ and appropriate p-value based on 1000 ancestral samples. When a test is significant, the corresponding cell entry is highlighted in bold. The letter next to the codon number represents consensus identification ('P' for positive)

		Counting Methods						Likelihood Methods			
Codon		*SLAC*		*WAC*		*Sampler*		*FEL*		*REL*	
35	P	**2.47**	(0.00)	**2.35**	(0.00)	**2.02:2.57**	(0.00:0.01)	**2.08**	(0.08)	**2.66**	(609915.00; 1.0000)
39		0.66	(0.14)	0.66	(0.14)	0.66:0.75	(0.11:0.14)	0.29	(0.59)	**0.02**	(997.53; 0.9896)
48		0.51	(0.25)	0.50	(0.25)	0.32:0.51	(0.25:0.39)	0.07	(0.91)	**0.02**	(570.98; 0.9820)
49		0.80	(0.17)	0.70	(0.21)	0.56:0.89	(0.14:0.29)	0.64	(0.46)	**0.02**	(68.78; 0.8677)
68		0.73	(0.11)	0.73	(0.11)	0.73:0.82	(0.09:0.11)	0.42	(0.46)	**0.02**	(225.75; 0.9556)
98		0.48	(0.29)	0.47	(0.30)	0.29:0.48	(0.29:0.41)	-0.38	(0.58)	**0.02**	(86.12; 0.8915)
102		0.57	(0.11)	0.57	(0.11)	0.57:0.57	(0.11:0.11)	**0.67**	(0.01)	0.02	(10.95; 0.5108)
123		0.40	(0.43)	0.24	(0.48)	-0.09:0.82	(0.31:0.60)	1.00	(0.39)	**0.83**	(4151.02; 0.9975)
135	P	**2.16**	(0.09)	**2.21**	(0.08)	1.85:2.33	(0.07:0.13)	1.80	(0.38)	**0.83**	(475.35; 0.9784)
165		0.74	(0.11)	0.74	(0.11)	0.74:0.74	(0.11:0.11)	0.36	(0.51)	**0.02**	(1331.89; 0.9922)
177	P	1.18	(0.16)	1.13	(0.18)	0.69:1.51	(0.09:0.31)	**1.65**	(0.10)	**0.83**	(507754.00; 1.0000)
178		1.80	(0.16)	1.72	(0.18)	1.50:1.87	(0.15:0.22)	1.70	(0.37)	**2.66**	(5830.06; 0.9982)
200	P	**3.42**	(0.00)	**3.33**	(0.00)	**3.23:3.58**	(0.00:0.00)	**2.53**	(0.02)	**2.66**	(6064080.00; 1.0000)
202	P	1.38	(0.16)	1.39	(0.16)	1.38:1.45	(0.15:0.16)	**1.56**	(0.10)	**0.02**	(117.80; 0.9183)
211	P	**3.01**	(0.01)	**2.76**	(0.01)	**2.80:3.20**	(0.00:0.01)	**4.47**	(0.00)	**0.83**	(1504750000.00; 1.0000)
215	P	**1.03**	(0.09)	0.93	(0.11)	0.84:1.13	(0.07:0.14)	**0.93**	(0.07)	**0.02**	(339017.00; 1.0000)

Table 2. Positively selected sites in HIV-1 Reverse Transcriptase of drug naïve patients identified by at least one of the methods. The first number for every method is an appropriately scaled $dN - dS$, so that they are directly comparable. The number in parentheses show p-values for the appropriate test, and the Bayes factor values for the REL method; posterior probabilities are also included for reference purposes, although they are not used in site classification. The entries for the Sampler method show 95% quantiles for the distribution of $dN - dS$ and appropriate p-value based on 1000 ancestral samples. When a test is significant, the corresponding cell entry is highlighted in bold. The letter next to the codon number represents consensus identification ('P' for positive)

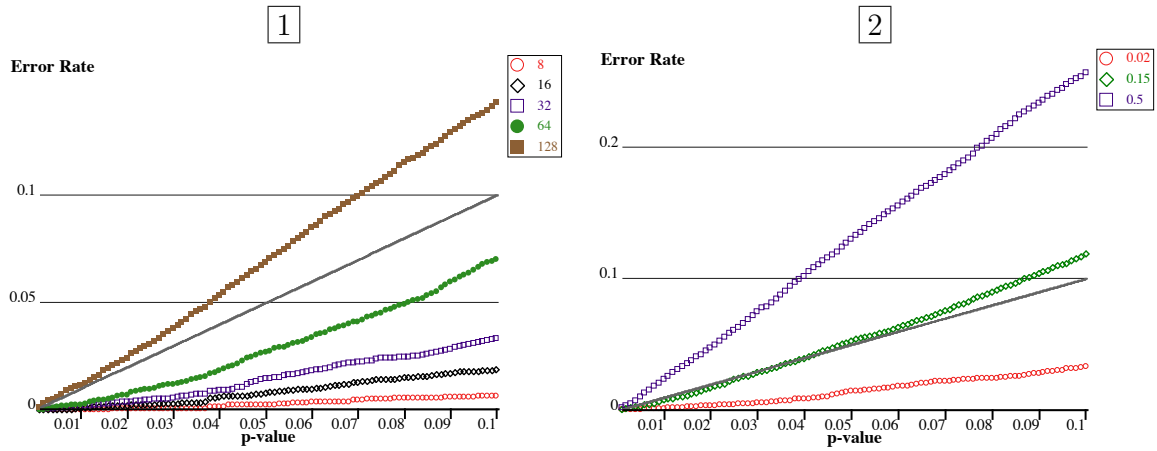

Figure 1. Type I errors at given critical levels (p-values) for neutrally evolving sequences analyzed with SLAC as functions of (1) number of sequences with average branch lengths of 0.02 and 250 codon alignments; and (2) average branch length of the tree, using 32 sequences and 250 codons per alignment. Solid gray line represents the expected error rate based on a given p-value; the predicted and observed error rates will coincide if the distribution of the test statistic used to assess significance (extended binomial) matches that derived from simulated data.

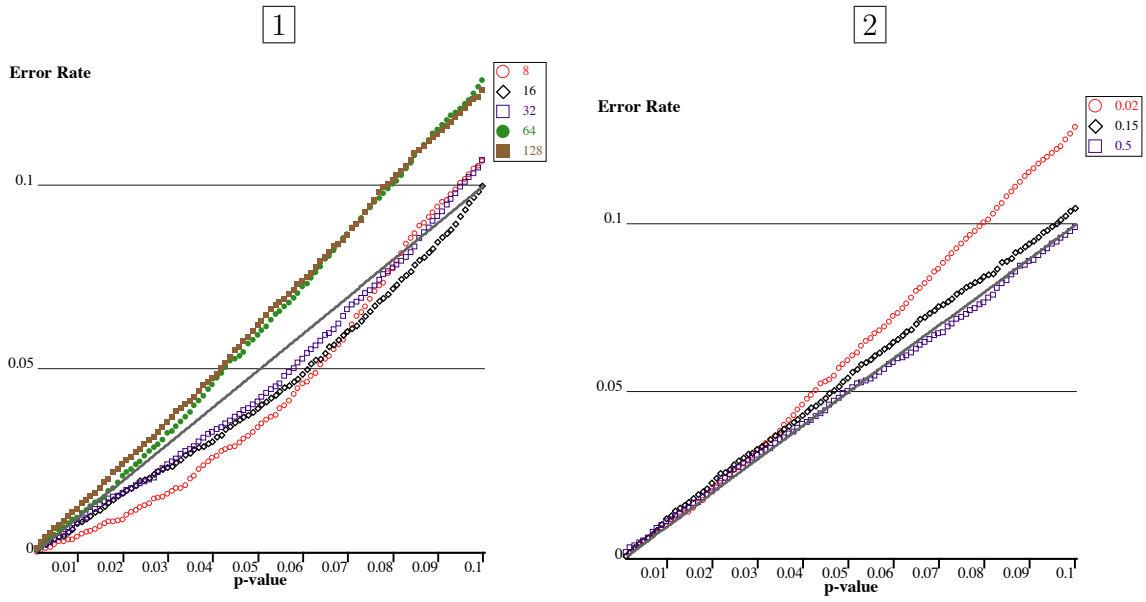

Figure 2. FEL Type I errors at given critical levels (p-values) for neutrally evolving sequences with 250 codons per alignment (1) as a function of the number of sequences with average branch length of 0.02; (2) as a function of average branch length using 64 sequences. Solid gray line represents the expected error rate based on a given p-value; the predicted and observed error rates will coincide if the distribution of the test statistic used to assess significance (χ^2_1) matches that derived from simulated data.

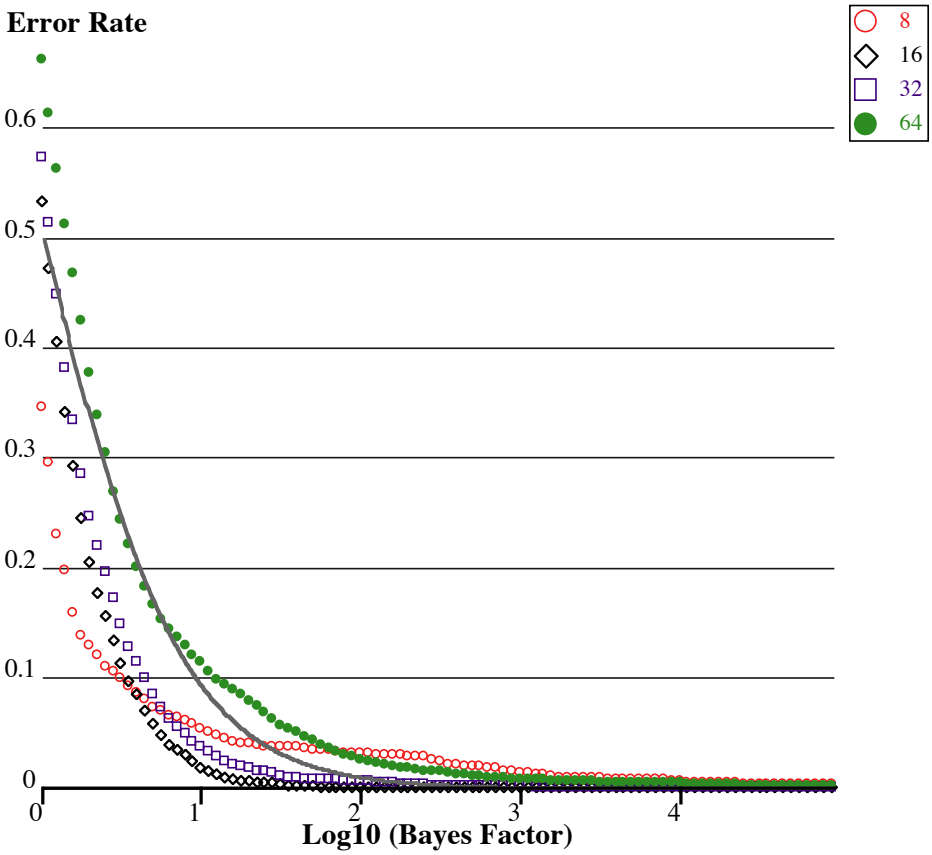

Figure 3. REL Type I errors at given critical levels (p-values) for neutrally evolving sequences with average branch lengths of 0.02 and 250−codon alignments as a function of the number of sequences. Solid gray line represents the expected error rate based on a given Bayes Factor with uninformative priors (so that prior odds are equal to 1).

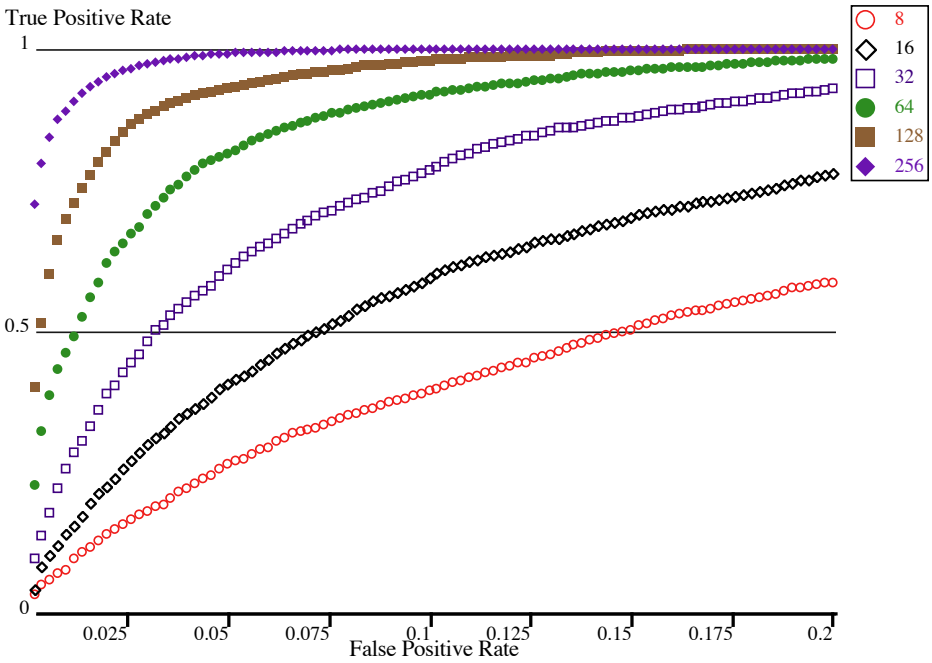

Figure 4. Perfomance of the SLAC method. 50 iterates evolved along symmetric trees with average branch lengths of 0.05 and 375 codons were used. Plots shown are ROC curves of the rates for detecting positively selected sites.

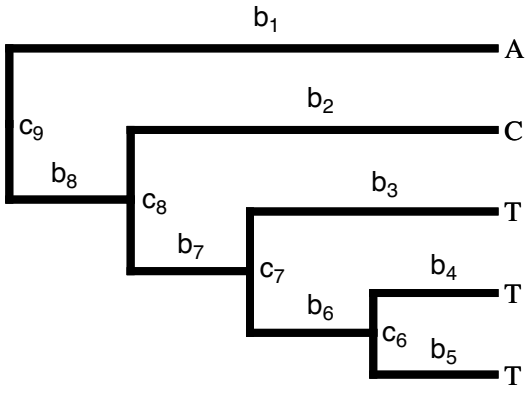

Figure 5. Example tree used for the WAC equivalence proof.

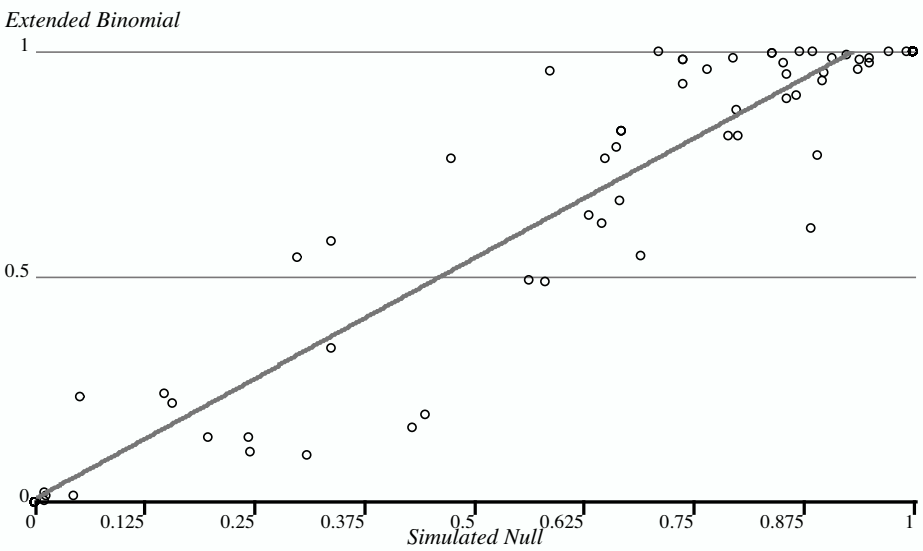

Figure 6. Correspondence between p-values for positive selection at a site using the extended binomial approximation and the simulated null distribution for the HIV-1 Reverse Transcriptase of drug naïve patients. There is a reasonably good linear relationship ($r^2 = 0.81$) between the two sets of p-values.

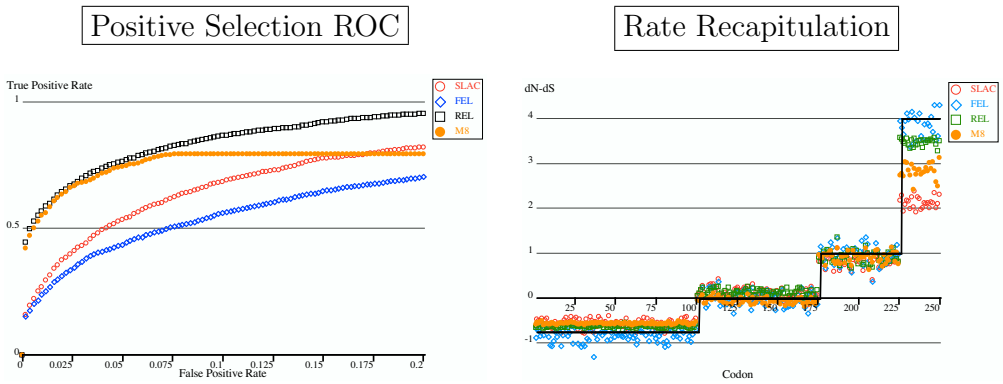

Figure 7. Receiver Operating Characteristic (ROC) curve mapping true positives versus false positives for detecting sites with $dN > dS$. Rate are recapitulated using average $dN - dS$ at a site inferred by each of the methods. The reference grey line represents the true values of $dN - dS$. Symmetric $32 - sequence$ tree with average branch length of 0.05 substitutions/site/unit time was used for data generation. 50 replicates were analyzed for each setting.