Molecular Biology and Evolution

The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection

William Fletcher[1,2] and Ziheng Yang*[,1,2]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[2]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London, United Kingdom

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Dan Graur

Abstract

The detection of positive Darwinian selection affecting protein-coding genes remains a topic of great interest and importance. The "branch-site" test is designed to detect localized episodic bouts of positive selection that affect only a few amino acid residues on particular lineages and has been shown to have reasonable power and low false-positive rates for a wide range of selection schemes. Previous simulations examining the performance of the test, however, were conducted under idealized conditions without insertions, deletions, or alignment errors. As the test is sometimes used to analyze divergent sequences, the impact of indels and alignment errors is a major concern. Here, we used a recently developed indel-simulation program to examine the false-positive rate and power of the branch-site test. We find that insertions and deletions do not cause excessive false positives if the alignment is correct, but alignment errors can lead to unacceptably high false positives. Of the alignment methods evaluated, PRANK consistently outperformed MUSCLE, MAFFT, and ClustalW, mostly because the latter programs tend to place nonhomologous codons (or amino acids) into the same column, producing shorter and less accurate alignments and giving the false impression that many amino acid substitutions have occurred at those sites. Our examination of two previous studies suggests that alignment errors may impact the analysis of mammalian and vertebrate genes by the branch-site test, and it is important to use reliable alignment methods.

Key words: indels, insertion, deletion, branch-site test, alignment, positive selection, codon models.

Research article

Introduction

The nonsynonymous to synonymous substitution rate ratio (ω) can be used to measure the selective pressure on the protein. A ratio $\omega < 1$ indicates purifying selection acting to preserve the amino acid sequence, whereas a neutrally evolving sequence will exhibit $\omega \approx 1$, and $\omega > 1$ represents positive selection driving the fixation of amino acid changes.

Many methods have been developed that aim to detect positive selection that affects specific lineages (Messier and Stewart 1997; Zhang and Kumar 1997; Yang 1998) or a subset of sites (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang et al. 2000), but both approaches may lack power. In the branch test, positive selection is detected on the branch only if ω averaged over all sites is significantly greater than 1, and similarly, the site test will detect positive selection only if the ω ratio averaged over all branches on the tree is greater than 1. As a result, both tests have generally been superseded by more powerful tests that are designed to detect episodic positive selection that affects only a few amino acid residues on a few lineages (Yang and Nielsen 2002; Guindon et al. 2004; Yang et al. 2005). The original branch-site test (Yang and Nielsen 2002) was found to generate excessive false positives when model assumptions were violated (Zhang 2004). However, a modified version (Yang et al. 2005) was found to have reasonable power and an acceptable false-positive rate under a variety of selection schemes (see Zhang et al. 2005 and below). This modified test has been widely used, for example, to investigate the adaptive evolution of genes underlying schizophrenia (Crespi et al. 2007) and possible positive selection affecting human disease genes (Vamathevan et al. 2008).

Although previous studies noted that different alignment methods may lead to different conclusions concerning detection of positively selected sites (Wong et al. 2008) and that alignment problems as well as poor sequence quality can cause spurious detection of positive selection by the branch-site test (Schneider et al. 2009; Mallick et al. 2010), the effects of insertions, deletions, and alignment errors on the branch-site test have not been systematically examined. In this paper, we use the recently developed simulation program INDELible (Fletcher and Yang 2009) to generate data sets under codon models incorporating indels to examine the performance of the test. The study is an update of Zhang et al. (2005). The effect of indels is examined by analysis of the true alignments and the effect of alignment errors by analysis of alignments generated using alignment programs, including PRANK (Löytynoja and Goldman 2005, 2008), MUSCLE (Edgar 2004), MAFFT (Katoh and Toh 2008), and ClustalW (Larkin et al. 2007).

Material and Methods

The Branch-Site Test of Positive Selection

We refer the reader to the original papers (Yang and Nielsen 2002; Yang et al. 2005) for further details of the

Table 1. Branch-Site Model.

Site Class	Proportion	Background	Foreground
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\ p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 \geq 1$
2b	$(1 - p_0 - p_1)\ p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$

Note.—This model is the alternative hypothesis for the branch-site test of positive selection. The null model is the same except $\omega_2 = 1$ is fixed.

branch-site test of positive selection. The model assumes that the branches on the phylogeny are divided a priori into foreground branches where some sites may be under positive selection and background branches where positive selection is absent. The model assumes four site classes (table 1). Site class 0 (with proportion p_0) includes codons that evolve under purifying selection on all lineages, with $0 < \omega_0 < 1$. Site class 1 (with proportion p_1) includes codons that evolve neutrally throughout the tree, with $\omega_1 = 1$. Codons in site classes 2a and 2b (with proportion $1 - p_0 - p_1$) are under positive selection on the foreground branches, with $\omega_2 > 1$, but are conserved or neutral on the background branches. The model involves four parameters in the ω distribution that are estimated from the data: p_0, p_1, ω_0, and ω_2. This branch-site model is the alternative hypothesis in the likelihood ratio test (LRT), and the null hypothesis is the same model but with $\omega_2 = 1$ fixed.

If the null hypothesis is true, twice the difference in log-likelihood between the two models ($2\Delta\ell$) should follow an asymptotic distribution that is a 50:50 mixture of point mass 0 and χ^2_1, with critical values of 2.71 and 5.41 at the 5% and 1% levels, respectively (e.g., Self and Liang 1987). We follow Zhang et al. (2005) and use χ^2_1 to conduct the test, with critical values of 3.84 and 5.99. This makes the test more conservative.

If the null hypothesis is rejected, a Bayes empirical Bayes (BEB) approach can be used to calculate the posterior probabilities that each site has evolved under positive selection on the foreground lineages (Yang et al. 2005).

Computer Simulation

INDELible (Fletcher and Yang 2009) was used to generate both the unaligned sequences and the true alignment. For easy comparison we followed Zhang et al. (2005) and used the two rooted trees (fig. 1). One branch on the tree is designated the foreground branch, whereas the others are the background branches. The transition/transversion rate ratio is fixed at $\kappa = 4$. Different from Zhang et al. (2005), the number of replicates used is 1,000 (instead of 200), the root sequence length is 300 codons (instead of 200), and the number of sites in each site class is random instead of being fixed. The stationary codon frequencies are those calculated from the base compositions at the three codon positions in a data set of five α and β mammalian globin gene sequences (data set abglobin.nuc in PAML; Yang 2007).

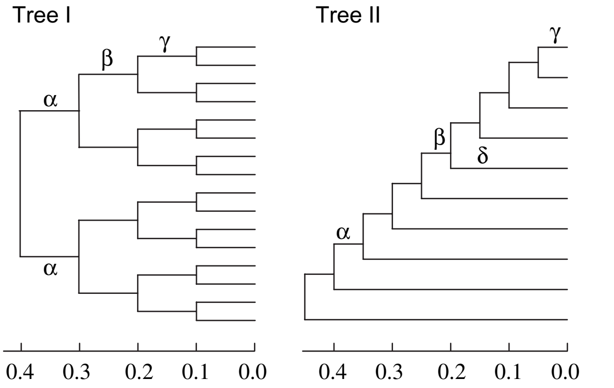

Fig. 1. Two model trees used in computer simulation. Branch lengths are drawn in scale, in terms of the number of synonymous substitutions per synonymous site. Greek letters indicate foreground branches used in the simulation.

As in Zhang et al. (2005), the simulation model assumes ten site classes. The foreground branch always uses selection scheme X, whereas the background branches use one of schemes X, Y, Z, U, or V (table 2). The ω values for site classes under the different selection schemes are listed in table 2. Schemes X, Y, and Z do not allow any sites under positive selection with $\omega > 1$, whereas schemes U and V do. Scheme X assumes some neutral sites (with $\omega = 1$), and other sites subject to varying degrees of negative selection (with $\omega < 1$). Scheme Y represents a partial relaxation of functional constraints where some sites have higher ω values than in scheme X. In scheme Z, all sites have $\omega = 1$, representing a complete relaxation of functional constraints. This is a very unrealistic scheme for any functional protein but is included partly because it may cause the test to generate false positives. In scheme U, some sites that experienced purifying selection in scheme X become positively selected, whereas scheme V differs from scheme X in a more complicated manner with some sites having lower ω and some having higher ω.

The molecular clock (rate constancy) holds for the synonymous substitution rate in both trees. In this study, as in Zhang et al. (2005), branch lengths are defined as the number of synonymous substitutions per synonymous site (d_S). For example, each branch in tree I represents about 10% of divergence at synonymous sites and, for the background

Table 2. The ω Values in Different Selection Schemes Used in Computer Simulation.

	Selection Schemes				
Site Class	X	Y	Z	U	V
1	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	0.70
3	0.80	1.00	1.00	4.00	4.00
4	0.80	0.90	1.00	0.80	0.80
5	0.50	1.00	1.00	2.00	2.00
6	0.50	0.75	1.00	0.50	0.50
7	0.20	1.00	1.00	0.20	0.30
8	0.20	0.60	1.00	0.20	0.20
9	0.00	1.00	1.00	0.00	0.10
10	0.00	0.50	1.00	0.00	0.00
Average ω	0.50	0.88	1.00	0.97	0.96

Note.—The proportion of each site class is 1/10.

scheme X with average $\omega = 0.5$ (table 2), 5% of divergence at nonsynonymous sites. However, INDELible defines branch lengths as the average number of substitutions per codon (t). They are related approximately by

$$t = 3(Nd_N + Sd_S) = 3d_S(\omega N + S)$$
$$\approx 3d_S(0.5 \times 0.7 + 0.3) = 1.95d_S$$

(Yang and Nielsen 2000), where N and S are the proportions of nonsynonymous and synonymous sites with $S \approx 0.3$ when $\kappa = 4$ (see fig. 3 of Yang and Nielsen 1998), and $\omega = 0.5$ is the average ω for the background scheme X. Therefore, for example, when a branch length of $d_S = 0.1$ is quoted in this paper, we have used $t = 0.195$ in INDELible. For simulations that included indels, the rates of insertion and deletion were set to be equal ($\lambda_I = \lambda_D$), and the ratio of substitutions to indels was similar to estimates in the literature, with $\lambda_S/(\lambda_I + \lambda_D) = 10$ (e.g., Ogurtsov et al. 2004). A geometric distribution was used to model insertion and deletion lengths, with parameter $q = 1 - p = 0.35$ chosen as it was deemed an adequate fit to published data on indels for protein-coding sequences in mammalian genomes (e.g., Taylor et al. 2004). The mean of this distribution is $1/p = 1.54$ codons, and the standard deviation is $\sqrt{q}/p{=}0.91$.

A random DNA sequence of 300 codons was generated at the root of the tree, by sampling from the stationary distribution and from the site classes specified in the ω scheme. The sequence is then "evolved" along the branches of the tree by simulating insertions and deletions, as well as substitutions. To preserve the reading frame, only insertions and deletions of whole codons are allowed. When new codons are inserted, they are assigned to one of the ten site classes at random. INDELible records the insertions and deletions that occurred on the tree, generating the sequences for the tips of the tree as well as the true alignment.

The sequences at the tips of the tree are aligned using the default options of the programs PRANK (version 081202; Löytynoja and Goldman 2005), MAFFT (version 6.716; Katoh and Toh 2008), MUSCLE (versions 3.7 and 4; Edgar 2004), and ClustalW (version 2.0.11; Larkin et al. 2007). The guide tree is calculated by those programs, and we did not provide the true tree to the alignment program as this option is not often available in real data analysis. To avoid out-of-frame indels, the codon sequences were translated into amino acid sequences, aligned, and then "back translated" into codon alignments. In addition, we used PRANK's "codon" option that uses the empirical codon model (ECM; Kosiol et al. 2007) to directly align the codon sequences although preserving the reading frame. This shall be referred to as PRANK (codon), with PRANK (aa) referring to the amino acid-based alignment.

The aligned codon sequences were analyzed using codeml from the PAML package (Yang 2007). Besides the estimated alignments, we also analyzed the true alignments. This allows us to evaluate the impact of insertions and deletions (in the true alignments) separately from that of alignment errors. Alignment gaps are either removed or kept. In the latter case, they are treated as missing data. The correct tree topology and correct identification of the foreground branches was assumed. The branch lengths on the unrooted tree were estimated by maximum likelihood without assuming the molecular clock. Then LRTs were performed at the 5% significance level, as described earlier. Analysis of each data set was conducted three times, with different initial values used for the numerical optimization to guard against codeml getting stuck in local maxima. In ~98% of data sets, all three analyses produced identical log-likelihood values, whereas in the remaining cases, the largest log-likelihood value was used in the LRT.

Measures of Alignment Quality

In order to investigate the effect of alignment errors on the branch-site test, we used two measures of alignment accuracy: the total column score (TC) and the sum of pairs score (SPS). TC is the proportion of columns from the true alignment that are reproduced exactly in the test alignment, and SPS is the proportion of aligned codon pairs from the true alignment that are also aligned together in the test alignment (Thompson et al. 1999). TC is more stringent than SPS. For the true alignment, TC = SPS = 1.

Results and Discussion

False-Positive Rate of the Branch-Site Test under Models of Relaxed Constraints

We investigated the false-positive rate of the test when the data were simulated using one of schemes X, Y, or Z as the foreground scheme (table 2). The background scheme was always X. Thus, the data were generated without positive selection, but the null hypothesis of the test is violated because the selection scheme is more complex than assumed by the null model. Note that the branch-site model is designed to test whether any residues in the protein experienced positive selection along the foreground branches (with $\omega > 1$) and is not intended to infer the detailed selection scheme on every branch or to estimate the ω ratio for every site and every branch. The latter task is hardly achievable due to lack of information in typical data sets (cf., Guindon et al. 2004).

We analyzed the data first with alignment gaps treated as missing data and then with gaps removed. The results are presented in table 3. The first set of analyses (column headed "No Indels") was conducted on data sets generated without indels in order to establish a baseline by which we could evaluate the effects of indels and alignment errors. The false-positive rate ranged from 0 to 4%, all below the nominal 5%. The results are similar to those of Zhang et al. (2005), even though there are differences in the two simulation experiments (in the number of codons, in codon usage frequencies, and in the use of fixed vs. random number of sites in each site class). A second set of analyses was conducted on the true alignments of data generated with indels (column headed "True Alignment"). The false-positive rate was also low, at 0–4%. Next, we analyzed the estimated alignments. The false-positive rates were 2–13% for PRANK (codon), 4–29% for PRANK (aa), 7–65% for MAFFT, 17–58% for MUSCLE v4, 9–73% for MUSCLE

Table 3. Frequencies of Cases in Which Positive Selection for Foreground Branches Is Erroneously Inferred by the Branch-Site Test (Type I Error).

					Gaps Kept							Gaps Removed						
Tree	FGB[a]	FGS[b]	ZNY 2005[c]	No Indels	True Alignment	PRANK (codon)	PRANK (aa)	MAFFT	Muscle v4	Muscle v3.7	ClustalW 2.0.11	True Alignment	PRANK (codon)	PRANK (aa)	MAFFT	Muscle v4	Muscle v3.7	ClustalW 2.0.11
I	α	X[d]	—	—	0.013	0.095	0.251	0.650	0.580	0.728	0.988	0.013	0.078	0.212	0.447	0.372	0.526	0.985
I	α	X	0.015	0.009	0.016	0.115	0.252	0.580	0.501	0.625	0.998	0.014	0.079	0.184	0.367	0.325	0.414	0.973
I	α	X[e]	—	—	0.009	0.130	0.288	0.414	0.410	0.473	0.943	0.006	0.075	0.193	0.275	0.270	0.307	0.851
I	β	X	0.020	0.012	0.016	0.106	0.179	0.325	0.320	0.383	0.841	0.014	0.077	0.144	0.236	0.227	0.247	0.670
I	γ	X	0.000	0.013	0.012	0.078	0.170	0.305	0.392	0.375	0.600	0.007	0.050	0.129	0.208	0.298	0.265	0.408
II	α	X	0.010	0.010	0.008	0.112	0.189	0.192	0.220	0.255	0.540	0.008	0.092	0.156	0.137	0.176	0.186	0.463
II	β	X	0.025	0.003	0.008	0.056	0.131	0.153	0.176	0.155	0.340	0.009	0.051	0.104	0.102	0.127	0.109	0.230
II	γ	X	0.020	0.008	0.007	0.021	0.036	0.075	0.279	0.093	0.118	0.010	0.024	0.029	0.058	0.243	0.077	0.100
II	δ	X	0.020	0.012	0.015	0.115	0.249	0.408	0.508	0.436	0.693	0.013	0.102	0.211	0.333	0.463	0.350	0.613
II	β[f]	X	0.030	0.008	0.015	0.076	0.119	0.149	0.175	0.168	0.340	0.010	0.054	0.094	0.092	0.118	0.107	0.241
II	β[g]	X	0.015	0.005	0.005	0.083	0.128	0.147	0.168	0.155	0.342	0.004	0.066	0.090	0.100	0.115	0.095	0.240
I	α	Y	0.030	0.019	0.018	0.045	0.103	0.407	0.361	0.493	0.999	0.020	0.049	0.096	0.255	0.230	0.319	0.989
I	β	Y	0.010	0.024	0.022	0.046	0.100	0.288	0.297	0.351	0.899	0.028	0.047	0.090	0.176	0.181	0.212	0.690
I	γ	Y	0.025	0.026	0.021	0.071	0.123	0.235	0.331	0.289	0.592	0.028	0.067	0.091	0.185	0.227	0.192	0.385
II	α	Y	0.020	0.016	0.016	0.104	0.254	0.256	0.248	0.302	0.685	0.016	0.078	0.194	0.159	0.176	0.192	0.580
II	β	Y	0.040	0.021	0.018	0.068	0.135	0.169	0.188	0.193	0.433	0.025	0.060	0.098	0.111	0.141	0.110	0.297
II	γ	Y	0.055	0.030	0.029	0.045	0.072	0.096	0.312	0.116	0.154	0.034	0.046	0.060	0.095	0.273	0.109	0.126
II	δ	Y	0.035	0.025	0.019	0.046	0.092	0.197	0.259	0.245	0.523	0.020	0.050	0.091	0.165	0.240	0.200	0.439
I	α	Z	0.025	0.030	0.022	0.033	0.064	0.388	0.322	0.486	0.999	0.036	0.043	0.080	0.246	0.221	0.333	0.978
I	β	Z	0.045	0.039	0.023	0.058	0.119	0.298	0.295	0.329	0.902	0.035	0.072	0.118	0.191	0.187	0.198	0.698
I	γ	Z	0.065	0.031	0.029	0.056	0.112	0.202	0.307	0.298	0.595	0.038	0.054	0.106	0.164	0.215	0.195	0.393
II	α	Z	0.010	0.014	0.011	0.129	0.251	0.236	0.242	0.331	0.702	0.017	0.074	0.199	0.149	0.181	0.213	0.600
II	β	Z	0.040	0.031	0.036	0.089	0.139	0.176	0.222	0.224	0.465	0.029	0.072	0.109	0.112	0.151	0.131	0.288
II	γ	Z	0.075	0.043	0.038	0.045	0.061	0.102	0.309	0.106	0.140	0.056	0.074	0.067	0.105	0.289	0.106	0.125
II	δ	Z	0.050	0.043	0.037	0.045	0.078	0.183	0.219	0.217	0.471	0.040	0.049	0.085	0.160	0.207	0.189	0.397

[a] FGB = foreground branch.

[b] FGS, foreground scheme. The background scheme is X in all cases. See table 2 for details of selection schemes X, U, and V.

[c] "ZNY 2005" refers to the results from Zhang et al. (2005).

[d] In these schemes there were insertions only (with no deletions).

[e] In these schemes there were deletions only (with no insertions).

[f] FGB β is not under positive selection (scheme X) but the two neighboring internal branches are both under positive selection with scheme V.

[g] FGB β is not under positive selection (scheme X) but all three neighboring branches (two internal and one terminal) are under positive selection with scheme V.

v3.7, and 12–100% for ClustalW. Finally, another set of analyses was performed on the same alignments after removing columns that contained gaps (using the cleandata option in codeml). With this approach, the false-positive rates were 0–6% for the true alignments and 2–10% for PRANK (codon), 3–21% for PRANK (aa), 6–45% for MAFFT, 11–46% for MUSCLE v4, 8–53% for MUSCLE v3.7, and 10–99% for ClustalW.

The results indicate that insertions and deletions do not cause the branch-site test to generate excessive false positives if the alignment is correct: The false-positive rates were at or below 5% in all cases and were very similar for the data generated with and without indels. However, false positives were often unacceptably high when the alignments were generated using the alignment programs. PRANK (codon) consistently produces the lowest false positives, PRANK (aa) tends to produce the second lowest false positives, with the two versions of MUSCLE and MAFFT coming in third, fourth, and fifth (none being clearly superior), and ClustalW generally performing worst. PRANK was the only method that had any false-positive rates below 5%. Compared with the true alignment, PRANK has room for improvement.

We conducted two simulations to investigate the robustness of the branch-site test to more complex variations in the ω ratio across lineages in the tree. We tested whether the branch-site test would be misled to produce false positives, in the presence of indels, when the foreground branch is not under positive selection but is surrounded in the phylogeny by background branches that are under positive selection. As in Zhang et al. (2005), we used tree II with foreground branch β and foreground scheme X. All other branches are background branches evolving under scheme X, except for the two internal branches adjacent to branch β that evolved under scheme V (table 2). In the second simulation, all three branches neighboring branch β evolved under scheme V. With the true alignment, the false-positive rate was ~0.01 in both simulations. As in Zhang et al. (2005), the branch-site test is robust and not misled by positive selection on branches close to the foreground branch of interest, even in the presence of indels and unequal codon frequencies.

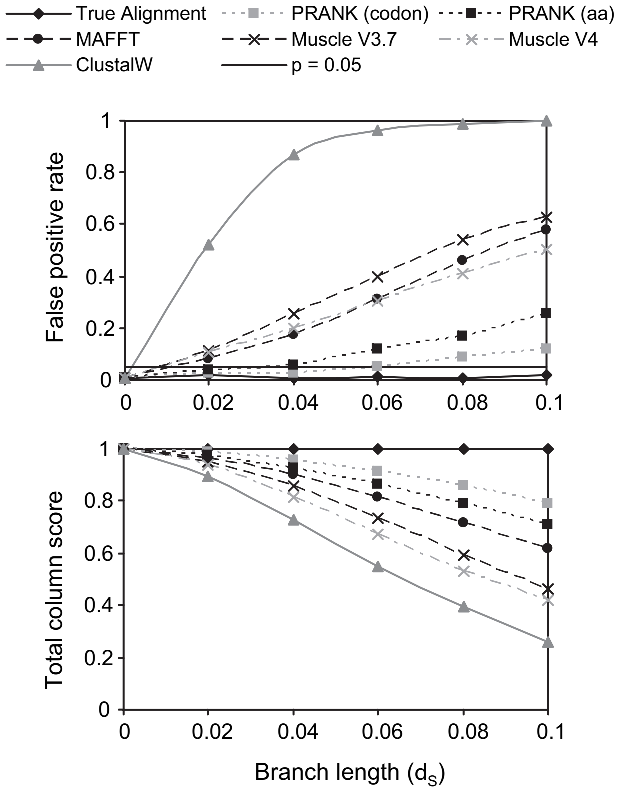

FIG. 2. False-positive rate (top) and alignment accuracy (TC) (bottom) for the different alignment methods, plotted against sequence divergence measured by the synonymous branch length (d_S) on tree I. Branch α in tree I was the foreground, with scheme X for all branches. Alignment gaps were treated as missing data.

Alignment quality is known to be closely related to sequence divergence. To investigate the effect of sequence divergence on the false-positive rate of the test, we kept the indel/substitution rate ratio constant and proportionally decreased the branch lengths in tree I to generate data at different divergence levels, using foreground branch α and foreground scheme X. Figure 2 shows the false-positive rate and alignment quality plotted against the sequence divergence, measured by the synonymous branch length d_S (note that all branches in tree I have the same length). The results when alignment gaps were removed were very similar and thus not shown. For all alignment methods, the false-positive rate decreased and alignment accuracy increased as sequences became less divergent. We also simulated data with insertions but no deletions and with deletions but no insertions. MUSCLE, MAFFT, and ClustalW were found to generate fewer false positives when there were deletions only but more false positives when there were insertions only (table 3). This result is consistent with the observation that a major problem with programs like ClustalW, MAFFT, and MUSCLE is that they do not deal with insertions properly, penalizing the same insertion event multiple times during the progressive alignment algorithm, although they deal with deletions more or less appropriately (Löytynoja and Goldman 2005). For PRANK, the pattern was the opposite. At any rate, whether there were deletions or insertions only, PRANK (codon) always had the lowest false positives, PRANK (aa) always came second, and ClustalW had the highest false positives, with MAFFT and the two versions of MUSCLE falling in-between.

Power of the Branch-Site Test in Detecting Positive Selection

To test how often the branch-site test correctly identifies positive selection on the foreground branches, we conducted simulations using either scheme U or V for the foreground branch. Scheme X is always used for the background branches. The results are shown in table 4. Again, use of the true alignments for data with indels produced very similar results to simulations without indels. For tree I, the power was 5–11% for data generated without indels and 4–11% for data with indels (table 4). Similarly, in tree II, the power was 1–32% without indels and 1–33%

Table 4. Frequencies of Cases in Which Positive Selection for Foreground Branches Is Correctly Inferred by the Branch-Site Test (Power).

					Gaps Kept							Gaps Removed						
Tree	FGB[a]	FGS[b]	ZNY 2005[c]	No Indels	True Alignment	PRANK (codon)	PRANK (aa)	MAFFT	Muscle v4	Muscle v3.7	ClustalW 2.0.11	True Alignment	PRANK (codon)	PRANK (aa)	MAFFT	Muscle v4	Muscle v3.7	ClustalW 2.0.11
I	α (0.2)	U	0.145	0.106	0.114	0.245	0.419	0.767	0.723	0.823	0.999	0.089	0.186	0.336	0.534	0.494	0.593	0.988
I	β (0.1)	U	0.110	0.055	0.042	0.146	0.270	0.451	0.441	0.492	0.892	0.042	0.107	0.189	0.289	0.269	0.300	0.718
I	γ (0.1)	U	0.095	0.054	0.061	0.165	0.285	0.452	0.538	0.523	0.750	0.058	0.128	0.225	0.317	0.376	0.354	0.525
II	α (0.05)	U	0.025	0.016	0.015	0.139	0.203	0.227	0.238	0.275	0.578	0.014	0.097	0.152	0.167	0.177	0.196	0.496
II	β (0.05)	U	0.040	0.026	0.022	0.111	0.171	0.189	0.208	0.197	0.394	0.024	0.081	0.135	0.117	0.156	0.120	0.295
II	γ (0.05)	U	0.095	0.062	0.065	0.102	0.126	0.175	0.379	0.202	0.231	0.049	0.064	0.101	0.138	0.311	0.163	0.171
II	δ (0.2)	U	0.220	0.182	0.188	0.359	0.530	0.659	0.749	0.721	0.871	0.134	0.257	0.427	0.556	0.662	0.600	0.783
II	β (0.15)	U	0.130	0.077	0.079	0.279	0.421	0.567	0.557	0.535	0.787	0.056	0.185	0.339	0.373	0.404	0.326	0.638
II	β (0.45)	U	0.180	0.192	0.149	0.670	0.851	0.946	0.925	0.937	0.997	0.124	0.492	0.744	0.845	0.816	0.844	0.990
II	γ and δ	U	0.330	0.309	0.331	0.471	0.590	0.681	0.774	0.714	0.864	0.258	0.323	0.455	0.555	0.680	0.566	0.753
I	α (0.2)	V	0.115	0.105	0.102	0.277	0.435	0.777	0.711	0.817	0.998	0.076	0.196	0.331	0.530	0.481	0.574	0.983
I	β (0.1)	V	0.075	0.063	0.052	0.165	0.292	0.486	0.470	0.516	0.918	0.046	0.109	0.217	0.299	0.288	0.320	0.732
I	γ (0.1)	V	0.075	0.071	0.066	0.181	0.270	0.433	0.530	0.504	0.731	0.055	0.112	0.180	0.297	0.381	0.337	0.515
II	α (0.05)	V	0.055	0.013	0.014	0.138	0.224	0.235	0.266	0.288	0.575	0.012	0.093	0.182	0.167	0.189	0.196	0.498
II	β (0.05)	V	0.035	0.027	0.031	0.101	0.164	0.200	0.241	0.204	0.417	0.028	0.072	0.122	0.122	0.162	0.116	0.292
II	γ (0.05)	V	0.120	0.062	0.079	0.100	0.123	0.171	0.396	0.208	0.222	0.066	0.082	0.098	0.131	0.329	0.148	0.171
II	δ (0.2)	V	0.300	0.174	0.181	0.364	0.527	0.676	0.761	0.733	0.896	0.139	0.261	0.429	0.569	0.697	0.612	0.809
II	β (0.15)	V	0.110	0.077	0.067	0.260	0.437	0.535	0.568	0.540	0.776	0.055	0.197	0.320	0.361	0.413	0.341	0.619
II	β (0.45)	V	0.190	0.154	0.147	0.615	0.831	0.949	0.928	0.936	0.997	0.119	0.462	0.725	0.821	0.808	0.803	0.985
II	γ and δ	V	0.435	0.321	0.298	0.441	0.547	0.638	0.764	0.687	0.831	0.224	0.307	0.410	0.520	0.677	0.540	0.733
II	β[d] (0.15)	V	0.195	0.159	0.169	0.619	0.800	0.902	0.878	0.867	0.963	0.114	0.427	0.694	0.748	0.745	0.677	0.902
II	β[e] (0.15)	V	0.500	0.226	0.251	0.438	0.588	0.672	0.687	0.657	0.877	0.173	0.311	0.468	0.487	0.498	0.458	0.729

[a] FGB = foreground branch. The length of the FGB (the number of synonymous substitutions per synonymous site) is shown in parentheses (see fig. 1).

[b] FGS = foreground scheme. The background scheme is X in all cases. See table 2 for details of selection schemes X, U and V.

[c] "ZNY 2005" refers to the results from Zhang et al. (2005).

[d] In this case, the root sequence was 900 codons in length instead of 300.

[e] In this case, ω ratios were doubled for classes 3 and 5 of scheme V.

with indels. Although these are relatively low detection rates, it should be noted that the average ω across all sites on the foreground branch under both schemes U and V is $<$1, and the average ω over all branches on the tree is never greater than 1 for any site class.

Next, we analyzed the alignments generated with the six alignment methods. The power was 10–67% for PRANK (codon), 12–85% for PRANK (aa), 17–95% for MAFFT, 21–93% for MUSCLE v4, 20–94% for MUSCLE v3.7, and 22–100% for ClustalW. If the gaps were removed, the power became 1–26% for the true alignments, 6–49% for PRANK (codon), 10–74% for PRANK (aa), 12–85% for MAFFT, 16–82% for MUSCLE v4, 12–84% for MUSCLE v3.7, and 17–99% for ClustalW. Thus, methods that had high false positives when there was no positive selection also had high true positives when there was positive selection. The conflicts between the accuracy and power of the test are considered in the next section, but in the rest of this section, we mainly focus on the true alignments and PRANK alignments.

We investigated the effect of sequence divergence on the power of the test. This was done in the same manner as in figure 2 except that we now use scheme U in place of scheme X for the foreground branch. The results are shown in figure S1 (Supplementary Material online). Power decreased and alignment accuracy increased as sequences became less divergent.

It may be expected that the method should be able to infer recent substitutions more reliably than ancient ones, and therefore it should be harder to detect positive selection on branches deeper in the phylogeny. This intuition appears to be correct for both schemes on both trees—in tree I the power is higher for branch γ than for branch β, and in tree II the power is higher for branch γ than for branch β, and higher for branch β than for branch α. However, the length of the foreground branch had a much greater effect—the branch-site test had greater power for branch α, the deepest branch in tree I, than for either of the shorter and more recent branches β and γ, and in tree II, the test had much greater power for the longer branch δ than for any of the shorter branches α, β, and γ.

To investigate the effect of foreground branch length, we performed further simulations with the length of branch β in tree II increased from 0.05 to 0.15 or 0.45. Under these conditions, the power of the test increased by 2- to 7-fold over the different alignments. When the foreground branch becomes too long, we would expect the power to decrease because of saturation of substitutions. Furthermore, we expect the power to increase when the same sites are under positive selection on several branches. To test this, we applied the same selection scheme to branches γ and δ on tree II and identified both as foreground branches when running codeml. Power was substantially higher than when γ or δ alone was the foreground branch (table 4) for both the true alignment and the PRANK (codon) alignment.

We also expect the power to be higher if the positive selection is stronger (with higher ω ratios) or if more sites are under positive selection. This was indeed the case. We conducted two sets of simulations using tree II, foreground branch β (with length 0.15), and foreground scheme V to examine such effects. In the first set, we increased the ω ratios for site classes 3 and 5 from 2 and 4 to 4 and 8, respectively (table 2). For the true alignment, power increased by more than 3-fold, whereas for the estimated alignments, the increase was 13–68%. In the second set of simulations, we increased the length of the root sequence from 300 to 900 codons, with three times as much data as before. For the true alignment, power increased more than 2-fold, with similar increases for PRANK.

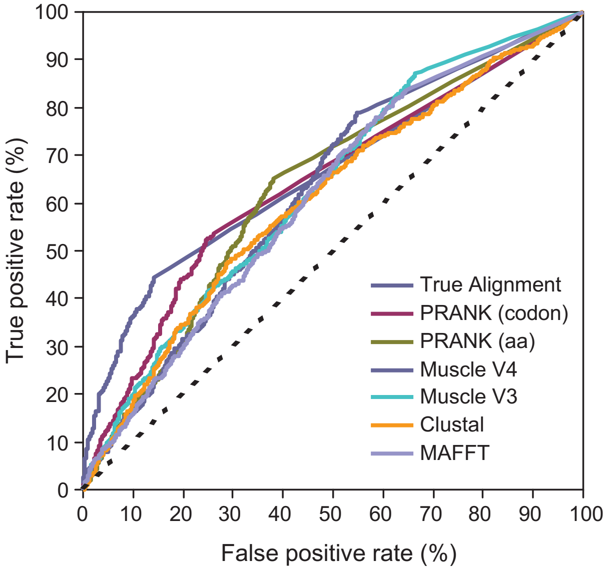

FIG. 3. ROC curves for the true alignment and the six alignment methods, when using foreground branch α and tree I. The false-positive rate is generated under foreground scheme X, and the true positive rate is generated under foreground scheme U.

Which Alignment Method is Best for Detecting Positive Selection?

A commonly used procedure for evaluating a test is the receiver operating characteristic (ROC) plot. An example is shown in figure 3 for simulations using branch α on tree I as the foreground branch. Scheme X was used for the foreground to calculate the false-positive rate, and U was used to calculate the true positive rate. Schemes X and U differ only in the positively selected site classes and can thus be used for a fair comparison. ROC curves that bulge toward the top left corner indicate a good predictor. It has been suggested that the area under the ROC curve (AUC) can be used as a summary to measure the performance of a method (Ling et al. 2003). Note that AUC = 0.5 for random guess and AUC = 1 for a "perfect" method that makes no mistakes. We calculated the AUC for each alignment method and foreground branch combination using the trapezoidal method of Hanley and McNeil (1983). The results are shown in table S1 (Supplementary Material online).

For the true alignment, AUC was smaller when gaps were removed than when they were kept, regardless of the tree or foreground branch. This is consistent with the expectation

Table 5. Alignment Accuracy for Different Alignment Methods

	Average Accuracies				Insertions or Deletions Only			
	TC		SPS		Insertions (Tree I)		Deletions (Tree I)	
	Tree I	Tree II	Tree I	Tree II	TC	SPS	TC	SPS
True alignment	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PRANK (codon)	0.79	0.63	0.95	0.91	0.83	0.96	0.73	0.94
PRANK (aa)	0.71	0.54	0.94	0.90	0.75	0.94	0.64	0.93
MAFFT	0.56	0.47	0.91	0.89	0.64	0.89	0.61	0.91
Muscle v4	0.42	0.44	0.88	0.86	0.42	0.87	0.50	0.90
Muscle v3.7	0.46	0.40	0.88	0.88	0.46	0.89	0.41	0.89
ClustalW 2.0.11	0.25	0.22	0.71	0.75	0.21	0.67	0.35	0.80

NOTE.— Simulations where root sequence length, branch lengths, or insertion/deletion rates were changed are excluded from the average accuracies.

that given the correctness of the alignment, removing gap columns amounts to reducing the amount of data available. For the estimated alignments, one aim of removing gaps is to reduce the alignment errors and the false positives of the test. However, it had this effect in only 6 of the 42 combinations of alignment method, tree, and foreground branch (twice for ClustalW and once for each of Muscle v4, MAFFT, and the two PRANK variations). Removing gaps before applying the branch-site test was thus ineffective in reducing alignment errors or false positives.

We thus focus on the left part of table S1 (Supplementary Material online), where gaps were kept. For tree I, the average AUC values were 0.63 for the true alignment, 0.62 for PRANK (codon), 0.61 for PRANK (aa), 0.60 for the two versions of Muscle, 0.59 for MAFFT, and 0.55 for ClustalW. For tree II, they were 0.63 for the true alignment, 0.63 for PRANK (codon), 0.61 for PRANK (aa), 0.60 for MAFFT and Muscle v3.7, 0.58 for Muscle v4, and 0.57 for ClustalW. The AUC values thus indicated that PRANK (codon) was the superior alignment method among those tested, followed by PRANK (aa), MAFFT, and Muscle, with ClustalW to be the poorest.

We consider the false detection of positive selection (false positive) to be a more serious error than a failure to detect positive selection (false negative) and suggest that a test with excessive false positives (with rate >20%, say) be avoided in real data analysis. With this viewpoint, the differences among the methods look even greater than suggested by the AUC values. For example, only PRANK had the false-positive rate under control in some data sets (table 3), but even PRANK leaves room for improvement. The rankings of the alignment methods are nevertheless the same whether we use the false positives or the AUC values.

Performance of BEB in Identifying Positively Selected Sites

When the branch-site test of positive selection is significant, the BEB procedure (Yang et al. 2005) can be used to calculate the posterior probability that a particular codon belongs to the class of positive selection. A codon with a high posterior probability is likely to have been evolving under positive selection on the foreground branches. We calculated the average frequency at which a codon was identified as being under positive selection using 95% or 99% cutoffs. We analyzed only the true alignments because alignment errors cause sites from different classes to be aligned together, making the calculations difficult. We only performed the BEB analysis on data sets in which the LRT was significant at the 5% level.

If BEB is conservative when evaluated under the Frequentist criterion (see Yang et al. 2005), one would expect the false positives to occur less than 5% of the time at the 95% cutoff and less than 1% of the time at the 99% cutoff. This was found to be true for data sets generated with indels, and with no positive selection (table 3), in 45 of 46 cases.

For the 23 cases where data sets were generated without indels, and under selection schemes X, Y, or Z, the false-positive rates were sometimes higher—at 6–7% for two cases at the 95% cutoff and 1–2% for five cases at the 99% cutoff. For all simulations with positive selection (table 4), the false-positive rate was very low (<0.1% at both cutoffs). However, the power of BEB in detecting positively selected codons was also very low, at ≤1% in all but one case.

Alignment Accuracy

The selection schemes and root sequence length had little effect on alignment accuracy. Thus, we averaged over the simulation conditions and present the results for each tree (table 5). The alignment accuracy is in the order PRANK (codon) > PRANK (aa) > MUSCLE v4 & MUSCLE v3.7 & MAFFT > ClustalW (table 5). This is the case for both trees and for all different simulation conditions (results not shown). This ranking was observed in ~100% (TC) and 88% (SPS) of the replicates for tree I and 87% (TC) and 43% (SPS) of the replicates for tree II. PRANK (codon) was clearly the best among the alignment methods examined here on most data sets. MAFFT was better than MUSCLE in most cases, but the relative performance of the two versions of MUSCLE was less clear. On average, TC judged MUSCLE v3.7 as better on tree I but worse on tree II, whereas SPS scored the two methods very similarly for both trees (table 5). It is noted that MUSCLE v4 is experimental. Overall, the order of alignment accuracy is exactly the opposite of the order for the false-positive rate discussed before.

To understand the nature of the alignment errors, we simulated data sets similarly to figure 2 but kept the substitution rate λ_S constant although increasing the indel rate $\lambda_I + \lambda_D$ (with $\lambda_I = \lambda_D$). We then calculated the average alignment length and the average number of distinct codons in a column after removal of columns with gaps. Programs MAFFT, MUSCLE, and ClustalW produced much shorter alignments than the true alignments, whereas PRANK alignments were of similar lengths (results not shown). The number of distinct codons in a column is shown in figure 4. This ranges from 1 to 16 for our data

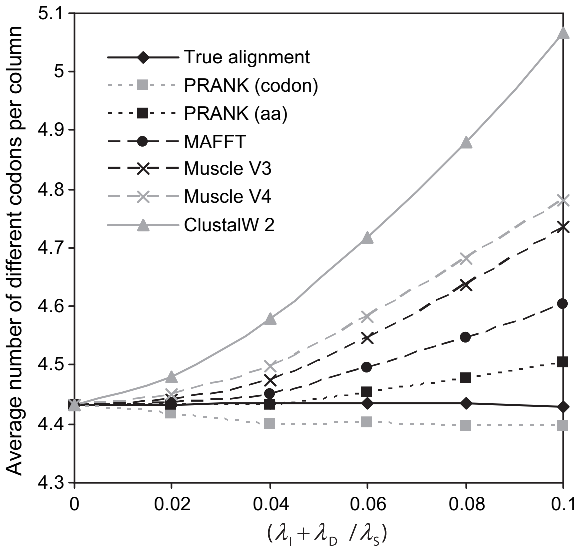

FIG. 4. Average number of different codons per column for different alignment methods, plotted against the indel/substitution rate ratio, in simulations with the substitution rate λ_S fixed. Tree I was used with scheme X for both foreground and background branches. Alignment gaps were removed.

with 16 sequences. For the true alignment, the number remained roughly constant regardless of the indel rates. For PRANK (codon) and PRANK (aa), this number is similar to that for the true alignment. For MAFFT, MUSCLE, and ClustalW, this number is much greater, especially at high indel rates. Those results are consistent with the observation of Löytynoja and Goldman (2005) that the main problem with those poor alignment methods is that they place nonhomologous codons (amino acids) into the same column. As such alignment errors remain after gaps are removed, the strategy of removing gaps to reduce the false-positive rate of the test is ineffective (table 3).

The nature of the alignment errors as discussed above suggests that the site models (Nielsen and Yang 1998; Yang et al. 2000) may be similarly affected by alignment errors, although site models are not examined in this paper. The impact of the alignment errors on the branch model (Yang 1998) appears to be more complex and may depend on the location of the foreground branch in the tree or whether errors are introduced when sequences on one side of the foreground branch are aligned against sequences on the other side.

The ECM (Kosiol et al. 2007) underlying PRANK (codon) was derived from the PANDIT database (Whelan et al. 2006), which has an average ω of 0.192 (Kosiol et al. 2007). Thus, PRANK (codon) should be more successful at aligning codons under selective constraint than those under positive selection. Similar bias may be expected for Prank (aa), MAFFT, MUSCLE, and ClustalW as they use empirical amino acid substitution/exchange matrices derived from large databases dominated by purifying selection. Furthermore, conserved amino acids correspond to less variable codons that are easy to align. In sum, codons under positive selection or under weak constraint are expected to be most prone to alignment errors. This prediction was found true for all six alignment methods. For example, figure 5 shows the alignment accuracy for codons in different site classes of scheme X for tree I for PRANK (codon). Codons in site classes with lower ω ratios were aligned more accurately. The background ω ratios had far greater effects on alignment quality than the foreground ω ratios. However, for a given site class with the same background ω ratio, alignment quality was slightly better for lower foreground ω ratios.

Recently, Hall (2008) suggested that a measure of "consistency" known as the Heads-or-Tails (HoT) score (Landan and Graur 2007) had a direct correlation with alignment accuracy and that it can be used to choose between alignments produced by different methods. The HoT score is the proportion of columns shared between the "Heads" alignment, generated from the original sequences, and the "Tails" alignment, generated from the reversed sequences. On our data, the HoT score chose MUSCLE v4 as the best method 93% of the time, whereas TC consistently favored PRANK (codon). This discrepancy appears to be due to the fact that PRANK breaks ties at random, whereas MUSCLE v4 makes the same choices so that alignment errors are consistent. We do not recommend the HoT score as a measure of alignment quality.

Implications for Past Studies of Positive Selection

What levels of sequence divergences may cause serious alignment errors and false detection of positive selection? To get a rough idea about this question, we examined two recent studies of positive selection using the branch-site test, one using five mammalian species (human, chimp, dog, mouse, and rat) (Vamathevan et al. 2008) and the other using a broader range of vertebrate species (five fishes, the *Xenopus* frog, the chicken, and at least four mammals) (Studer et al. 2008). Both studies used MUSCLE v3 to construct the alignments and both removed columns with gaps before applying the branch-site test. Although, Studer et al. realigned some of their genes using MAFFT and obtained highly similar results, we note that consistency between MAFFT and MUSCLE was not a good indication for high alignment quality in our simulations.

The MUSCLE alignments for the 3,081 mammalian genes were provided by Vamathevan J (personal communication). The gene sequences were realigned using PRANK (codon), and both sets of alignments were analyzed using the branch-site test, with the human or chimpanzee lineages designated as the foreground branch. Following Vamathevan et al. (2008), we used the 5% significance level and the Bonferronni correction for multiple testing (Anisimova and Yang 2007). The initial analysis of Vamathevan et al. identified 69 (2.2%) and 354 (11.5%) genes under positive selection on the human and chimpanzee lineages, respectively (Vamathevan 2008, table 3.1). The counts from our reanalysis of the same data were nearly identical, at 70 and 355. Our analysis of the PRANK (codon) alignments produced 33 (1.1%) and 83 (2.7%)

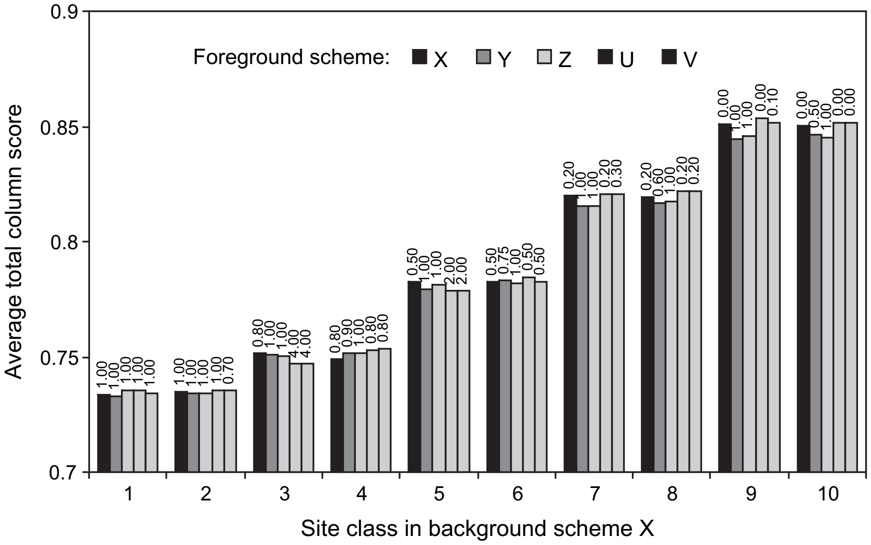

Fig. 5. Average alignment accuracy for PRANK (codon) for codons in the ten site classes of the background scheme X and for different foreground ω ratios. Tree I is used. The number above each column is the foreground ω ratio for that site class.

genes for the human and chimpanzee branches, respectively, much smaller than for the MUSCLE alignments, indicating that the original MUSCLE alignments may involve substantial alignment errors. Problems with the MUSCLE alignments were noted by Vamathevan et al., who applied a manual curation step and reported 54 (1.8%) and 162 (5.3%) positively selected genes for the human and chimpanzee branches, respectively (Vamathevan et al. 2008, table 1). These counts are much smaller than from the original MUSCLE alignments but are still much higher than those from the PRANK (codon) alignments.

The MUSCLE alignments for the 767 vertebrate genes were downloaded from http://bioinfo.unil.ch/supdata/Singleton.html (Studer et al. 2008). Both the original alignments and the PRANK (codon) realignments were analyzed using the branch-site test, with three foreground branches considered: the mammal lineage, the euteleosts lineage, or the bony vertebrate lineage. Studer et al. (2008) used the 1% significance level and identified 8%, 25%, and 31% of genes to be under positive selection along the three branches. Our reanalysis of the same original alignments produced 8%, 24%, and 30%. The counts were 6%, 16%, and 18% from analysis of the PRANK (codon) alignments.

Those comparisons suggest that mammalian and vertebrate gene sequences are divergent enough for the impact of alignment errors on the branch-site test to be a real concern. This conclusion is consistent with the results of Schneider et al. (2009) and Mallick et al. (2010). Many past studies detecting positive selection in divergent genes may benefit from a reanalysis using alignments generated from a more reliable method such as PRANK (codon).

Conclusion

In this paper, we investigated the accuracy and robustness of the branch-site test in the presence of insertions, deletions, and alignment errors. Our results obtained from analyses of the true alignments suggested that indels have little effect on the performance of the branch-site test. In the presence of indels, the test is still robust to violation of model assumptions such as the existence of more than three site classes, more than one site class evolving under positive selection, or more than one kind of background branches. The BEB method for detecting positively selected sites was noted to have low false positives under such conditions.

Our study highlighted the importance of alignment quality to detection of positive selection using the branch-site test. In particular, most current alignment programs tend to place nonhomologous codons (or amino acids) in the same column, misleading the test into claiming excessive amino acid changes at those sites. Removing alignment gaps helped to reduce false positives only slightly. We found that PRANK was superior to MAFFT, MUSCLE, and ClustalW. In particular, PRANK (codon) produced the most accurate alignments, with the lowest false-positive rates. Nevertheless, even PRANK (codon) does not have the false-positive rate under control. It is hard to imagine tests of positive selection that are tolerant of gross alignment errors, and we suggest that it may be more profitable to try and improve current alignment algorithms.

Supplementary Material

Figure S1 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This study was conducted using high-performance computing cluster LEGION of University College London. W.F. is supported by an Engineering and Physical Sciences Research Council/Medical Research Council Doctoral Training Centre studentship. Z.Y. is a Royal Society Wolfson Merit Award holder.

References

Anisimova A, Yang Z. 2007. Multiple hypothesis testing to detect adaptive protein evolution affecting individual branches and sites. *Mol Biol Evol*. 24:1219–1228.

Crespi B, Summers K, Dorus S. 2007. Adaptive evolution of genes underlying schizophrenia. *Proc Biol Sci*. 274:2801–2810.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32: 1792–1797.

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

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*. 101:12957–12962.

Hall BG. 2008. How well does the HoT score reflect sequence alignment accuracy? *Mol Biol Evol*. 25:1576–1580.

Hanley JA, McNeil BJ. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148:839–843.

Katoh K, Toh H. 2008. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*. 9:212.

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

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*. 24:1380–1383.

Larkin MA, Blackshields G, Brown NP, et al. (11 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.

Ling CX, Huang J, Zhang H. 2003. AUC: a better measure than accuracy in comparing learning algorithms. Advances in artificial intelligence. Springer: Berlin (Germany). p. 329–341.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A*. 102:10557–10562.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320:1632–1635.

Mallick S, Gnerre S, Muller P, Reich D. 2010. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*. 19:922–933.

Messier W, Stewart C-B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–154.

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

Ogurtsov AY, Sunyaev S, Kondrashov AS. 2004. Indel-based evolutionary distance and mouse-human divergence. *Genome Res*. 14:1610–1616.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol*. 2009:114–118.

Self SG, Liang K-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc*. 82:605–610.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res*. 18:1393–1402.

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

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res*. 14:555–566.

Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*. 27:2682–2690.

Vamathevan J. 2008. Evolutionary analysis of mammalian genomes and associations to human disease [PhD thesis]. [London]: University College London.

Vamathevan J, Hasan S, Emes R, et al. (11 co-authors). 2008. The role of positive selection in determining the molecular cause of species differences in disease. *BMC Evol Biol*. 8:273.

Whelan S, de Bakker PI, Quevillon E, Rodriguez N, Goldman N. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res*. 34:D327–D331.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473–476.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15: 568–573.

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

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

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*. 17:32–43.

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

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

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

Zhang J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*. 21: 1332–1339.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol*. 14: 527–536.

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