Erasing Errors due to Alignment Ambiguity When Estimating Positive Selection

Benjamin Redelings*,1,2
1Biology Department, Duke University
2The National Evolutionary Synthesis Center, Durham, NC
*Corresponding author: E-mail: benjamin.redelings@duke.edu.
Associate editor: Tal Pupko

Abstract

Current estimates of diversifying positive selection rely on first having an accurate multiple sequence alignment. Simulation studies have shown that under biologically plausible conditions, relying on a single estimate of the alignment from commonly used alignment software can lead to unacceptably high false-positive rates in detecting diversifying positive selection. We present a novel statistical method that eliminates excess false positives resulting from alignment error by jointly estimating the degree of positive selection and the alignment under an evolutionary model. Our model treats both substitutions and insertions/deletions as sequence changes on a tree and allows site heterogeneity in the substitution process. We conduct inference starting from unaligned sequence data by integrating over all alignments. This approach naturally accounts for ambiguous alignments without requiring ambiguously aligned sites to be identified and removed prior to analysis. We take a Bayesian approach and conduct inference using Markov chain Monte Carlo to integrate over all alignments on a fixed evolutionary tree topology. We introduce a Bayesian version of the branch-site test and assess the evidence for positive selection using Bayes factors. We compare two models of differing dimensionality using a simple alternative to reversible-jump methods. We also describe a more accurate method of estimating the Bayes factor using Rao-Blackwellization. We then show using simulated data that jointly estimating the alignment and the presence of positive selection solves the problem with excessive false positives from erroneous alignments and has nearly the same power to detect positive selection as when the true alignment is known. We also show that samples taken from the posterior alignment distribution using the software BA1-Phy have substantially lower alignment error compared with MUSCLE, MAFFT, PRANK, and FSA alignments.

Key words: sequence alignment, Bayes factor, positive selection, false-positive rate, insertion/deletion, codon models.

Introduction

Phylogenetic methods are an essential tool for inferring biological properties of nucleotide sites using evolutionary models. Phylogenetic methods make use of homologous sequence data from multiple species to infer site properties from the patterns of nucleotide differences between species. Such properties may include the presence or absence of functional constraint (Siepel et al. 2005), the presence of diversifying positive selection (Goldman and Yang 1994; Muse and Gaut 1994), and the ability of DNA sites to bind particular proteins (Sinha et al. 2004). All phylogenetic methods for inferring site properties share in common the reliance on a phylogenetic tree (known or estimated) and a multiple sequence alignment. The multiple sequence alignment is essential for inferring site properties because it specifies which nucleotides from different sequences are homologous, and therefore what counts as a “site.” Current methods for inferring site properties rely on a previously computed estimate of the alignment. Errors in the alignment may therefore lead to the estimation of spurious properties for sites that are incorrectly aligned and for the sequence as a whole.

Alignment error is especially problematic when estimating diversifying positive selection, because aligning nonhomologous residues is likely to imply a spurious nonsynonymous substitution, which will then be interpreted as evidence for positive selection. Alignment errors, together with sequencing errors and the inclusion of nonhomologous genes and exons, substantially raise the frequency of erroneously detecting positive selection. Alignment errors have therefore limited the utility of phylogenetic site-annotation methods in practice (Schneider et al. 2009; Villanueva-Cañas et al. 2013). For example, in whole-genome comparative analyses of yeast (Wong et al. 2008) and of Drosophila (Markova-Raina and Petrov 2011), the choice of alignment program had a large effect on which genes were identified as experiencing positive selection. Furthermore, the majority of positives in these whole-genome studies were actually false positives arising from misaligned codons. Simulation studies show that errors in estimated multiple sequence alignments can lead to substantially inflated false-positive rates (FPRs) in inferring positive selection (Fletcher and Yang 2010), even in methods that have quite conservative FPRs when the true alignment is known.

To mitigate this problem, researchers have searched for alignment methods with the lowest error rates in detecting positive selection (Jordan and Goldman 2012; Privman et al. 2012). These studies found PRANK (Löytynoja and Goldman 2005) alignments to be superior to alignments from MUSCLE.
If \( \omega > 1 \), then we say that the codon is undergoing diversifying positive selection. Note that diversifying positive selection is therefore a preference for amino acid change per se.

To use such models to assign properties to individual codon sites in a gene, Nielsen and Yang (1998) introduced models in which different codon sites may choose from a fixed collection of \( \omega \) values. These \( \omega \) values, and the fraction of sites that evolve according to each one, are themselves unknown parameters to be estimated from data. Thus, for example, in the M2a model (Wong et al. 2004), some fraction \( p_0 \) of sites have \( \omega = \omega_0 \leq 1 \), some fraction \( p_1 \) have \( \omega = 1 \), and the remainder have \( \omega = \omega_2 \geq 1 \) (table 1). One can obtain a model without positive selection by constraining \( \omega_2 = 1 \) and \( 1 - p_0 - p_1 = 0 \), and this leads to an LRT for positive selection (Nielsen and Yang 1998). This test assesses the evidence that there are any sites that are positively selected and thus does not require the external correction for multiple testing that would be needed if each site was tested separately. (Wong et al. 2004). However, note that if we mistakenly align two separate subcolumns into a single (incorrect) column, then we may create a spurious nonsynonymous substitution. Because even a single column undergoing positive selection is considered a rejection of the null hypothesis of no positive selection, it is possible that this test may be sensitive to alignment error.

Zhang et al. (2005) describe an extension of this model that allows positive selection to be both site specific and branch specific. The tree topology is fixed and assumed to be known a priori, as are the branches on which positive selection might occur. These branches are labeled “foreground” branches, whereas the remainder are labeled “background” branches. In this “branch-site” model, the \( \omega \) for a site may switch to \( \omega_2 \geq 1 \) on the foreground branches, whereas remaining either \( \omega_0 \leq 1 \) or \( \omega_1 \geq 1 \) on all background branches (table 2). Some fraction \( p_2 \) of conserved sites undergo this switch; the same fraction of neutral sites undergo this switch. Zhang et al. (2005) suggest constructing an LRT by comparing this model with a null model where \( \omega_2 \) is constrained to 1. This null model is preferred over the null model where \( 1 - p_0 - p_1 \) is constrained to 0, because it allows \( \omega \) to change to 1 on the foreground branches even when there is no positive selection. This avoids treating relaxation of selective constraints as positive selection and avoids false positives.

Table 1. The M2a Model for Site-Dependent \( \omega \).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>( 0 \leq \omega_0 \leq 1 )</td>
<td>( \omega_1 = 1 )</td>
<td>( \omega_2 \geq 1 )</td>
</tr>
<tr>
<td>Frequency</td>
<td>( p_0 )</td>
<td>( p_1 )</td>
<td>( 1 - p_0 - p_1 )</td>
</tr>
</tbody>
</table>

Table 2. The Branch-Site Model for Branch- and Site-Dependent \( \omega \).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>( \omega_0 )</td>
<td>1</td>
<td>( \omega_0 )</td>
<td>1</td>
</tr>
<tr>
<td>Foreground</td>
<td>( \omega_0 )</td>
<td>1</td>
<td>( \omega_2 )</td>
<td>( \omega_2 )</td>
</tr>
<tr>
<td>Frequency</td>
<td>( p_0 )</td>
<td>( p_1 )</td>
<td>( p_2 \omega )</td>
<td>( p_2 \omega )</td>
</tr>
</tbody>
</table>
when the data do not follow the simple model used in inference (Zhang 2004).

**Positive Selection and Alignment Uncertainty**

Fletcher and Yang (2010) showed that on data sets simulated under a variety of evolutionary scenarios containing insertions and deletions, alignment errors can lead to FPRs for the branch-site test that are substantially higher than 0.05. Despite the fact that these simulation models contradict the assumptions of the branch-site model used for inference, Fletcher and Yang (2010) found no evidence of excessive false positives when the true alignment was used. In contrast, when estimated alignments were used, the FPRs depended strongly on the evolutionary scenario that was simulated and on the software program used to reconstruct the alignment. Under some evolutionary scenarios, the use of alignments from ClustalW led to FPRs as high as 0.99. Other alignment software performed better, with PRANK codon-based alignments having the lowest FPRs. Nevertheless, PRANK alignments had FPRs as high as 0.13 under some evolutionary scenarios, substantially exceeding the desired level of 0.05.

Fletcher and Yang (2010) simulated from models that extend the branch-site model by allowing each column to select from a variety of different neutral or conservative \( \omega \) values on background branches, following Zhang (2004) and Zhang et al. (2005). Table 3 presents one of these simulation models, which we term FY+. The FY+ model expands the number of site classes from 4 to 10. This allows it to express the standard branch-site model. For example, in the FY+ model, \( \omega \) may take on the values 0.0, 0.2, 0.5, 0.8, and 1.0 on background branches, instead of just 0.0 and 1.0. More importantly, neutral \( \omega \) values and some conserved \( \omega \) values never switch to a new \( \omega \) value on the foreground branch under the FY+ model. This contradicts the standard branch-site model, which assumes that each neutral or conserved category of \( \omega \) has the same probability \( p_2 \) of switching to a new \( \omega \) value on foreground branches. These simulation conditions are useful for determining how the branch-site test behaves under model violations. However, such model violations may decrease ability to detect positive selection.

In this article, we use the FY+ simulation conditions from Fletcher and Yang (2010) to assess the effect of alignment error on inference of positive selection under the branch-site model. However, because these simulation conditions violate the assumptions of the branch-site model, we have introduced simulation conditions BS1 and BS2 to assess performance when the model is not violated. The FY, BS1, and BS2 simulation conditions are described in detail in the Materials and Methods.

### Joint Estimation

Integrating over all alignments under an evolutionary model is a more natural approach to estimation under alignment uncertainty (Thorne and Kishino 1992; Allison and Wallace 1994). Instead of censoring parts of the alignment that are difficult to align, multiple alternative alignments are considered with an appropriate weight that depends on the data and the evolutionary model. This approach is a more natural evolutionary approach to the problem of alignment uncertainty because it treats insertions and deletions as mutations occurring on particular branches of a phylogenetic tree, instead of merely gaps in a matrix. Alignment estimation therefore benefits from the use of an evolutionary model that includes the phylogeny and thus should achieve greater accuracy than heuristic alignment programs that do not have access to the evolutionary tree (Löytynoja and Goldman 2005).

Integrating over all alignments is statistically more natural because it conducts inference starting from the observed data, which are unaligned sequences. A multiple sequence alignment estimate \( \hat{A} \) is not observed and so is not considered data. This affects the likelihood, because the likelihood is defined to be proportional to the probability of the data, given hypothesis \( H \) and parameters \( \Theta \). The likelihood does not condition on \( \hat{A} \), because \( \hat{A} \) is not a model parameter.

\[
\Pr(\text{unaligned data} \mid H, \Theta) \neq \Pr(\text{unaligned data} \mid H, \Theta, \hat{A}).
\]

Instead, the likelihood integrates over the alignment \( \hat{A} \), because \( \hat{A} \) is a latent variable:

\[
\Pr(\text{unaligned data} \mid H, \Theta) = \sum_A \Pr(\text{unaligned data}, \hat{A} \mid H, \Theta).
\]

Support for positive selection might then be phrased in terms of a ratio of marginal likelihoods or Bayes factor:

\[
BF_{10} = \frac{\sum_A \Pr(\text{unaligned data}, \hat{A} \mid H = 1)}{\sum_A \Pr(\text{unaligned data}, \hat{A} \mid H = 0)}.
\]

Here, \( H = 0 \) indicates that the null model (\( H_0 \), no positive selection) is true, whereas \( H = 1 \) indicates that the alternative model (\( H_1 \), with positive selection) is true. To perform model selection between the \( H_0 \) and \( H_1 \) models, we can incorporate both of these models into a larger probability expression:

\[
\Pr(\text{unaligned data}, A, \Theta, H).
\]

We can then perform MCMC to estimate the posterior probability (PP) that \( H = 1 \), which we can use to compute the Bayes factor.

### Table 3. The FY+ Simulation Conditions for Branch- and Site-Dependent \( \omega \).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>Class 7</th>
<th>Class 8</th>
<th>Class 9</th>
<th>Class 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Foreground</td>
<td>1.0</td>
<td>1.0</td>
<td>4.0</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Incorporating alignment estimation inside the test in this way allows joint estimation of the alignment and the presence of positive selection. This is important because it allows each of $H_0$ and $H_1$ to be evaluated in the context of alignments that are adapted to that model, instead of evaluating both models on a common alignment estimate $\tilde{\mathbf{A}}$. In such an approach, not only does the alignment influence estimates of positive selection but the two models of selection (with and without positive selection) also influence the alignment. We note that the ratio of marginal likelihoods could in theory be replaced with a ratio of maximum likelihoods to allow the construction of a LRT that incorporates alignment uncertainty.

Instead of censoring an alignment estimate to remove ambiguous regions, we therefore propose to remove excess false positives by jointly estimating the alignment and the presence of positive selection. We do this by integrating over near-optimal alignments inside the test for positive selection. We introduce a Bayesian version of the branch-site test recommended by Zhang et al. (2005). The combination of the $H_0$ and $H_1$ substitution models can be referred to as the branch-site testing (BST) model. We then extend the software program BAli-Phy (Redelings and Suchard 2005) to perform this test while integrating over all alignments under the RS07 insertion/deletion model (Redelings and Suchard 2007). The full model may then be referred to as the BST/RS07 model. This approach allows alignment estimation to achieve greater accuracy by allowing site-dependent conservation heterogeneity for both the $H_0$ and $H_1$ substitution models. The approach incorporates multiple different sources of alignment uncertainty, including alignment uncertainty due to uncertainty in insertion/deletion parameter values, and alignment uncertainty due to near-optimal alignments.

**Bayesian Model Selection**

We perform model selection in the Bayesian framework based on the Bayes factor (Jeffreys 1998; Suchard et al. 2001). The Bayes factor for a model is an odds ratio that quantifies the strength of evidence for (or against) that model in terms of the relative fit of the data to each model. Bayes factors above 20:1 are often considered “strong” support, whereas Bayes factors above 3:1 but less than 20:1 are considered “positive” support, and Bayes factors less than 3 are “not worth more than a bare mention” (Kass and Raftery 1995). We compute $BF_{H_p}$, which is the Bayes factor in favor of $H_1$ against $H_0$, and thus quantifies the evidence in favor of positive selection. To compute Bayes factors, we must supply prior distributions on unknown variables in the model. There is no explicit penalty for higher-dimensional models in the Bayesian framework. Instead, higher-dimensional models suffer an implicit penalty when the prior distribution on the additional dimensions does not focus all of its mass on the value that happens to have the highest likelihood. The most influential prior distributions for the branch-site model are the prior distribution on $p_1$ and the prior distribution on $\omega_2$. These priors play a crucial role in defining $H_1$ because they determine what fraction $p_1$ of sites display positive selection under $H_1$, as well as the strength $\omega_2$ of positive selection. Other prior distributions are less influential because they are shared by both $H_0$ and $H_1$ and are likely to affect both models equally.

Although Bayes factors can be computed without specifying the prior probability that $H = 0$ and $H = 1$, PPs cannot. We choose to set the prior probability of $H_0$ and $H_1$ to 0.5. This prior distribution on $H$ treats both models equally a priori and corresponds to the assumption that 50% of genes experience positive selection and 50% do not. When we compute the false-discovery rate (FDR), we also make the assumption that the ratio of genes with and without positive selection is 1:1 and refer to the result as FDR_{1:1}. In this “equipoise” scenario, the Bayes factor equals the posterior odds. Posterior probabilities > 0.952 then correspond to a Bayes factor > 20:1, whereas PPs > 0.75 correspond to a Bayes factor > 3:1.

In the rest of this article, we first describe how to estimate the Bayes factor with sufficient accuracy. We then simulate data according to the FY scenario examined by Fletcher and Yang (2010) and proceed to test the accuracy of joint estimation of alignment and positive selection. We compare the FPR, TPR, and FDR of joint inference with inference based on the known true alignment and with inference based on the use of a single fixed alignment estimate from MUSCLE, PRANK, FSA, or MAFFT. We then simulate additional data sets according to the BS1 and BS2 scenarios. We compare the FPR, TPR, and FDR of the traditional branch-site LRT and the Bayesian version of the branch-site test. We also compare the accuracy of alignments from MUSCLE, PRANK, FSA, and MAFFT to alignments sampled from the posterior alignment distribution under the BST/RS07 model using BAli-Phy.

**Results**

**Improved Estimator for the Posterior Odds of Positive Selection**

As described above, we introduce a variable $H$ to indicate which model is in effect. When $H = 1$, the likelihood is computed under the positive selection model, and when $H = 0$, the likelihood is computed under the model without positive selection. Under this scheme, the probability of positive selection is $Pr(H_1 \mid \text{data}) = Pr(H = 1 \mid \text{data})$, and the probability of no positive selection is $Pr(H_0 \mid \text{data}) = Pr(H = 0 \mid \text{data})$.

At each iteration of the Markov chain, a new value of $H$ is sampled, because $H$ is part of the state of the Markov chain. Let us define $h_j$ to be the value of $H$ sampled at the $j$th iteration of the Markov chain. Here, $h_j$ will be 0 or 1. The usual way of estimating $Pr(H = 1 \mid \text{data})$ from N MCMC samples would simply be to compute the fraction of samples in which $H = 1$:

$$Pr(H = 1 \mid \text{data}) \approx \frac{1}{N} \sum_{j=1}^{N} 1[h_j=1].$$ (1)

Here, we use the mathematical notation $1_{\{ \cdot \}}$, which is defined to be 1 if the condition $\{ \cdot \}$ is true, and 0 otherwise. This method of estimating $Pr(H = 1 \mid \text{data})$ does not work very well if the probability is near 1 or 0. Suppose we have $N = 100$ samples. In that case, it is not possible to obtain a probability between 99/100 and 100/100. Although these two PPs may
seem similar, they lead to the very different odds ratios of 99/1 and 100/0 = ∞. The posterior odds is closely related to the Bayes factor and thus to the strength of evidence for positive selection. We seek an estimator for the posterior odds that can attain values between N − 1 and ∞. This is necessary to compute high posterior odds without obtaining an enormous number of samples from the Markov chain.

Our strategy for obtaining improved estimates of \( Pr(H = 1 \mid \text{data}) \) is to record at each iteration not only the value of \( H \) but also the expected value of \( H \) given all other variables in the Markov chain. To show that the expected value can be used to construct a valid estimator for \( Pr(H = 1 \mid \text{data}) \), we define \( X \) to refer to all variables in the Markov chain except \( H \). We note that

\[
Pr(H = 1 \mid \text{data}) = E \left( Pr(H = 1 \mid X) \mid \text{data} \right) = E \left[ E (1_{H=1}) \mid X \right] \mid \text{data} = E [Pr(H = 1 \mid X) \mid \text{data}],
\]

where the inner expectation is over \( H \) and the outer expectation is over \( X \). Now let \( x_j \) be the value of \( X \) sampled at the \( j \)th iteration of the Markov chain. Then the approximation

\[
Pr(H = 1 \mid \text{data}) \approx \frac{1}{N} \sum_{j=1}^{N} Pr(H = 1 \mid X = x_j, \text{data})
\]

allows us to approximate \( Pr(H = 1 \mid \text{data}) \) by averaging over the value of \( Pr(H = 1 \mid X = x_j, \text{data}) \) that is recorded at each iteration.

To compute \( Pr(H = 1 \mid X = x_j, \text{data}) \), we modify the software to compute \( Pr(H = 0, X = x_j, \text{data}) \) and \( Pr(H = 1, X = x_j, \text{data}) \) at each iteration without changing \( h_j \). Then

\[
Pr(H = 1 \mid X = x_j, \text{data}) = \frac{Pr(H = 1, X = x_j, \text{data})}{Pr(X = x_j, \text{data})} = \frac{Pr(H = 1, X = x_j, \text{data})}{Pr(H = 0, X = x_j, \text{data}) + Pr(H = 1, X = x_j, \text{data})}.
\]

Taking the conditional expectation of an estimator to obtain an improved estimator, as we have done here, is sometimes called Rao-Blackwellization because a similar process is described in the Rao–Blackwell theorem (Blackwell 1947). This theorem also guarantees that the new estimator (eq. 2) has a variance that is at least as small as the variance of the old estimator (eq. 1) and is frequently smaller. We note that the new estimator allows estimates of posterior odds between \( N − 1 \) and \( ∞ \).

### How PPs Change with Different Alignments

After simulating 1,000 data sets with diversifying positive selection (FY +) throughout the entire gene region and 1,000 data sets without positive selection (FY −), we performed the Bayesian version of the branch-site test on each data set using a variety of alignment methods. The effect of alignment error on the PP of positive selection can be illustrated by plotting the PP given the true alignment against the PP for various alignment estimation methods (fig. 1). In such a plot, each point represents a simulated data set. These plots show that when MUSCLE or MAFFT alignments are used, the PP of positive selection is increased for nearly all data sets, and the increase is frequently large. When FSA or PRANK alignments are used, for many data sets, the PP is similar to the PP from the true alignments. However, when the PP is different, it is usually an increase, and the increases may be small or large. FSA seems to experience larger increases of PP than PRANK. In contrast, when jointly estimating alignments, decreases in PP seem to be as frequent as increases, and the magnitudes are not large. When fixing a single alignment sampled from the posterior distribution of the coestimation analysis, PPs are nearly as accurate as when performing a full coestimation analysis, at least under these simulation conditions. We further illustrate the effect of alignment error on PPs by plotting the distribution of PPs across data sets for each alignment method (fig. 2). For MUSCLE, MAFFT, FSA, and PRANK alignments, PPs are shifted toward 1.0. However, PPs under joint estimation have nearly the same distribution as when the true alignment is known.

We calculate the squared correlation across data sets of PP for estimated alignments versus PP for true alignments. For data sets simulated under the FY + model without positive selection, the squared correlation coefficient is 0.088 for MUSCLE, 0.080 for MAFFT, 0.17 for FSA, 0.26 for PRANK, 0.66 when fixing a posterior sampled alignment, and 0.79 when coestimating the alignment. Squared correlations on data sets simulated under the FY + model with positive selection are 0.10 for MUSCLE, 0.15 for MAFFT, 0.37 for FSA, 0.49 for PRANK, 0.75 when fixing a posterior sampled alignment, and 0.85 for jointly estimating alignments.

### Discriminating between Data Sets with and without Positive Selection

To assess the ability of different methods to discriminate between data sets with and without positive selection, we computed ROC curves for Bayesian inference of positive selection using different alignment methods (fig. 3). ROC curves allow comparison of different methods at the same level of FPR, even if those methods achieve different FPR values in practice. Depending on the method, the alignment was either coestimated (Joint As) under the BST/RS07 model or fixed to an externally supplied alignment estimate. We supplied external estimates from the alignment reconstruction programs MUSCLE, MAFFT, FSA, and PRANK. We additionally supplied the known true alignment (True A) and a single fixed alignment (Joint A) sampled from the posterior distribution of the coestimation analysis. The TPR and FPR were computed based on the FY + and FY − data sets, respectively (table 4). At an FPR of 5%, Bayesian inference based on fixing the true alignment attains a TPR of 30% (True A). Jointly estimating the alignment yields a TPR of 27% (Joint As), whereas fixing a posterior sampled alignment yields a power of 25% (Joint A). In contrast, conditioning on alignments estimated by PRANK, FSA, MUSCLE, or MAFFT leads to a TPR of 15%, 15%, 11%, and 9%, respectively. Thus, use of
the true alignment allows substantially better ability to discriminate between data sets with and without positive selection than use of alignment estimates from PRANK, FSA, MUSCLE, or MAFFT. Joint estimation achieves nearly the same ability to discriminate as when the true alignment is known.

Joint Estimation Avoids Inflated FPRs
We computed the FPR for Bayesian inference of positive selection using different alignment methods. We report the FPR for both the BF > 3:1 criterion and the BF > 20:1 criterion in table 4. For the “True A,” “Joint A,” and “Joint A” analyses, Bayesian tests based on the BF > 3:1 and BF > 20:1 criterion...
all yielded an FPR of \(<1\%\). Under the BF \(>3:1\) criterion, estimating alignments with PRANK, FSA, MAFFT, or MUSCLE lead to FPRs of 6\%, 13\%, 43\%, and 52\%, respectively. Under the BF \(>20:1\) criterion, estimating alignments with PRANK, FSA, MAFFT, or MUSCLE lead to FPRs of 6\%, 13\%, 43\%, and 52\%, respectively. Thus, the use of estimated alignments leads to inflated FPRs for the Bayesian tests relative to use of the true alignment. However, joint estimation did not lead to inflated FPRs.

### Performance of the LRT and Bayesian Branch-Site Tests

The Bayesian tests and the branch-site LRT have similar trade-offs between FPR and TPR on the FY, BS1, and BS2 simulated data sets as illustrated by their ROC curves (fig. 4). For comparisons between the Bayesian and LRT approaches, we assume that the true alignment is known to focus on the difference in approach. The BS1 simulation conditions yield little power to detect positive selection, the BS2 simulation conditions lead to higher power, and the FY simulation conditions are intermediate. Under the FY and BS1 simulation conditions, the Bayesian and LRT approaches lead to nearly identical ROC curves. However, under the BS2 simulation conditions, Bayesian inference leads to a ROC curve that clearly dominates the LRT curve. For example, at an FPR of 5\%, the LRT has a TPR of 59\%, whereas Bayesian inference has a TPR of 76\%.

Although the ROC curves for the LRT and Bayesian tests are similar, the Bayesian tests tend to select more conservative points on these curves that have lower FPR, TPR, and FDR (table 5). For example, on the FY data sets, the standard branch-site LRT based on the conservative \(\chi^2_1\) distribution has a 1\% FPR, a 13\% TPR, and an 8\% FDR\(_{1:1}\). (Use of the true asymptotic distribution \(\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1\) would lead to a 2\% FPR, a 20\% TPR, and a 10\% FDR\(_{1:1}\).) In contrast, use of the BF \(>3:1\) criterion for the Bayesian test leads to an FPR of \(<1\%\), a TPR of 7\%, and an FDR\(_{1:1}\) of 3\%, whereas the use of the BF \(>20:1\) criterion leads to an FPR of \(<1\%\), a TPR of \(<1\%\), and an FDR that is unknown because the FPR and TPR are too small.

Because the FY simulation conditions violate the assumptions of the branch-site model, we also examined performance under the BS1 and BS2 simulation conditions, which do not violate the assumptions. For the BS1 data sets, the branch-site LRT achieves a 2\% FPR, a 3\% TPR, and a 36\% FDR\(_{1:1}\). The BF \(>3:1\) criterion for the Bayesian test attains an FPR \(<1\%\), a 3\% TPR, and a 31\% FDR, whereas the BF \(>20:1\) criterion attains an FPR and TPR that are both \(<1\%\) and an unknown FDR. On the BS2 data sets, the branch-site LRT achieves a 4\% FPR, a 53\% TPR, and a 6\% FDR\(_{1:1}\). Bayesian inference under a BF \(>3:1\) criterion attains a 9\% FPR, an 84\% TPR, and a 10\% FDR\(_{1:1}\); under the BF \(>20:1\) criterion it attains a \(<1\%\) FPR, a 43\% TPR, and a \(<1\%\) FDR\(_{1:1}\). Note that the \(P < 0.05\) criterion for the LRT and the BF \(>3:1\) criterion for the Bayesian test both achieve FPR \(<5\%\) under the BS1 simulation conditions but still achieve an FDR\(_{1:1}\) \(>30\%\).

### Table 4. Performance of Bayesian Tests under Different Alignment Estimates.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Attribute</th>
<th>True A (%)</th>
<th>Joint As (%)</th>
<th>Joint A (%)</th>
<th>PRANK (%)</th>
<th>FSA (%)</th>
<th>MAFFT (%)</th>
<th>MUSCLE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPR = 5%</td>
<td>FPR</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>TPR</td>
<td>30</td>
<td>27</td>
<td>25</td>
<td>15</td>
<td>15</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>FDR(_{1:1})</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>25</td>
<td>25</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>FPR = 1%</td>
<td>FPR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TPR</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>FDR(_{1:1})</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>12</td>
<td>20</td>
<td>29</td>
<td>38</td>
</tr>
<tr>
<td>BF &gt; 3:1</td>
<td>FPR</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>6</td>
<td>13</td>
<td>43</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>TPR</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>18</td>
<td>31</td>
<td>63</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>FDR(_{1:1})</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>24</td>
<td>29</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td>BF &gt; 20:1</td>
<td>FPR</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>3</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>TPR</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>(&lt;1)</td>
<td>4</td>
<td>10</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>FDR(_{1:1})</td>
<td></td>
<td></td>
<td></td>
<td>?</td>
<td>?</td>
<td>26</td>
<td>36</td>
</tr>
</tbody>
</table>

**Fig. 3.** ROC curves for inferring positive selection using different methods of alignment. Vertical dotted line indicates 5\% FPR. Diagonal dotted line describes the performance of a random guess at different levels of specificity. Each curve represents Bayesian estimation based on alignments estimated using MUSCLE, MAFFT, FSA, PRANK, a sampled alignment (Joint A), or joint estimation averaging over alignments (Joint As), as indicated.
Measuring Alignment Error

Sampling from the posterior alignment distribution under the BST/RS07 model yields alignments with less pairwise alignment error than alignments taken from MUSCLE, MAFFT, FSA, or PRANK. We examined the relationship of pairwise alignment error versus the evolutionary distance for each alignment method. Each multiple alignment contains a large number of pairwise alignments, because it implies a pairwise alignment between each pair of sequences at the tips of the tree. The evolutionary distance between tips in the tree in Figure 5 can only be 0.2, 0.4, 0.6, or 0.8. Figure 6 plots the pairwise alignment error versus evolutionary distance for alignments estimated using MUSCLE, MAFFT, FSA, PRANK (DNA), PRANK (codon), and PRANK (aa) and by sampling an alignment from the posterior alignment distribution (Joint A). The pairwise alignment error appears to be approximately linear as a function of evolutionary distance between the two sequences, and so we report alignment error at an evolutionary distance of 0.8 as a representative measurement. MUSCLE has the highest amount of alignment error of the methods we tested, with an average alignment error of 0.178. MAFFT is similar, with an alignment error of 0.143. FSA has an alignment error of 0.103. The PRANK variants all perform similarly, with an average alignment error of 0.077 (DNA), 0.086 (codon), and 0.098 (aa). Sampling from the posterior alignment distribution yields the smallest error, with an average alignment error of 0.042.

We also explored the differences in alignments produced by different alignment methods by measuring the tendency of each method to produce alignments longer or shorter than the known true alignments in the FY/C0 data sets. Figure 7 shows the joint distribution of the true alignment length and estimated alignment length for MUSCLE, MAFFT, FSA,
Erasing Errors due to Alignment Ambiguity · doi:10.1093/molbev/msu174

PRANK, and alignments sampled from the posterior alignment distribution under the BST/RS07 model. We also computed the median difference for each method between the estimated alignment length and the true alignment length. A score of 0 would indicate that the method is just as likely to overestimate the length as to underestimate it. **MUSCLE**, **MAFFT**, and **FSA** tend to underestimate the true alignment length, with median differences of –49, –39, and –30 codons, respectively. **PRANK** alignments had a median difference of +16 codons, whereas posterior sampled alignments had a median difference of +1 codon. Thus, alignments sampled under the BST/RS07 model are nearly unbiased, whereas other methods tend to be biased upward or downward. To provide a scale, the median length of true alignments was 434 codons. We also note that, under the BST/RS07 model, the 95% credible interval for alignment length has a mean width of 11.1 codons across data sets, whereas the 50% credible interval has a mean width of 4.75 codons. Thus, uncertainty in alignment length under the BST/RS07 model is smaller than the biases of other reconstruction methods.

**Discussion**

Our study indicates that jointly inferring the alignment and the presence or absence of positive selection eliminates the problem of high FPR for detecting diversifying positive selection from estimated alignments. This is partly due to increased accuracy in alignment estimation under the BST/RS07 model. We find that alignments sampled from the posterior distribution have a pairwise alignment error that is about half that obtained by **PRANK**, which is one of the best aligners to use when detecting positive selection (Fletcher and Yang 2010). Posterior sampled alignment lengths were also more accurate than alignment lengths estimated using **MUSCLE**, **MAFFT**, **FSA**, and **PRANK**. Use of alignments sampled from the posterior under the BST/RS07 model successfully eliminated inflated FPRs, as other alignment estimates could not. However, the ability to integrate over alignment uncertainty provided a small but measurable increase in accuracy for detecting positive selection and gave PPs of positive selection that were more similar to PPs given the true alignment. **PRANK** alignments, while not as accurate as posterior sampled alignments, were substantially more accurate than **MUSCLE** and **MAFFT** alignments and lead to more accurate inferences of positive selection. **FSA** alignments were nearly as accurate as **PRANK** alignments but yielded slightly worse FPRs in estimating positive selection. Despite their similar performance in detecting positive selection, **FSA** and **PRANK** alignments have different characteristics, because **FSA** alignments tend to be shorter than the true alignment, whereas **PRANK** alignments tend to be longer.

**Alignment Uncertainty**

Our approach to integrating out alignment uncertainty takes into account many sources of alignment uncertainty that may be divided into two categories: Parameter uncertainty and near-optimal alignments. First, uncertainty in evolutionary process parameters can cause alignment uncertainty when plausible changes to these parameters lead to different alignment estimates. Evolutionary process parameters include branch lengths, gap parameters such as the insertion and deletion rate, substitution parameters such as transition and transversion rates, and the evolutionary tree. Second, even when the evolutionary process parameters are fully known, there may be thousands of alignments that achieve an optimal or nearly optimal probability. It is not possible to choose a single alignment from this cloud of possibilities without discarding many plausible alternatives. To fully account for alignment uncertainty, a procedure must account for both near-optimal alignments and the effect of uncertain parameters on the alignment.

The ability to account for both parameter uncertainty and near-optimal alignments is a natural feature of Bayesian inference, which handles uncertainty from both latent variables (such as the alignment) and from parameters (such as indel rates). This differs from a number of current alignment-cen-tering methods, which usually consider only one source of alignment uncertainty. For example, **GUIDANCE** (Penn, Privman, Ashkenazy, et al. 2010; Penn, Privman, Landan, et al. 2010) considers uncertainty in the phylogeny but does not explicitly consider uncertainty due to near-optimal alignments or due to uncertainty in other parameters such as gap penalties. **HoT** (Landan and Graur 2008) considers uncertainty due to near-optimal or co-optimal alignments but does not consider uncertainty resulting from uncertainty in parameters such as the phylogeny or gap penalties.

In this article, we have simulated sequences on a fixed tree and assumed that the tree topology was known. As a result, there is no uncertainty about the evolutionary tree topology and thus no alignment uncertainty that could result from tree topology uncertainty. This assumption is probably adequate in some scenarios such as the Drosophila 12 genomes project (Markova-Raina and Petrov 2011). However, in other cases it
is inadequate, either because the topology is the primary focus of estimation or because the unknown topology is a nuisance parameter. In such a case, it is possible that GUIDANCE could perform better because it explores a source of uncertainty that is not considered here. To incorporate uncertainty resulting from the unknown tree, we could simply enable the topology-sampling MCMC moves already present in BAi-Phy (Redelings and Suchard 2005, 2007). However, the branch-site model prevents this, because it requires the researcher to label the foreground branches a priori. These branches must then be known to be part of the true tree a priori, and thus, the topology must be constrained before the estimation is begun. A model such as that proposed by Pond et al. (2011) would solve this problem by allowing the set of branches experiencing positive selection to be coestimated along with the alignment. Alternatively, researchers could simply switch to a model that has site specific but not branch-specific effects, like the M2a model and the M8a model. These models are already available in BAi-Phy.

Mismatch between Models and Reality

This study examines the effect of alignment error on inferring positive selection by primarily examining the FY simulation conditions. However, many biological sequences do not match these simulation conditions. For example, envelope gene sequences from HIV contain regions with a much higher insertion/deletion rate (Privman et al. 2012). Because we do not examine such conditions in the article, it remains an open question how well our method would perform on such data sets.

This simulation study also does not address a number of ways in which real data could violate assumptions made in the RS07 insertion/deletion model used here. For example, indel lengths in nature probably do not follow a geometric length distribution (Cartwright 2006), and they can sometimes occur within codons instead of between them (Redelings and Suchard 2007). The rates of insertions and deletions may frequently depend on which letters that are inserted or deleted. For example, tandem-repeat indels have a higher rate than other indels (Golenberg et al. 1993). More importantly, different regions of a DNA sequence may have substantially different insertion–deletion rates. By forcing a single sequence-wide indel rate, the insertion/deletion model in this article will of necessity underestimate indel rates in indel hot spots and overestimate indel rates in cold spots. In such cases, we expect that the power and accuracy of alignment integration will lag behind knowledge of the true alignment more substantially than it does in this article. Simulation studies such as Privman et al. (2012) that include variation of rates over different regions may be able to reveal how much power is lost.

Bayesian Formulation of the Branch-Site Test

Here, we have focused on inferring positive selection for a single gene using Bayes factors. We assumed that it was equally likely for a gene to be with and without positive selection. An alternative approach would be to infer the prior probability \( \pi_1 \) that a gene contains positive selection by analyzing many genes simultaneously. For example, if there were \( G \) different genes and gene \( g \) has model \( H_g \), we could use the following hierarchical prior:

\[
\pi_1 \sim \text{Uniform}(0,1) \\
H_g \sim \text{Bernoulli}(\pi_1) \quad \text{for } g = 1 \ldots G
\]

Such an approach would not be computationally prohibitive, because it is possible to do inference by computing the Bayes factor for each gene separately and then combining the Bayes factors. For data sets containing many genes, we recommend such an approach, because it would naturally require stronger evidence to infer positive selection when the fraction of genes experiencing positive selection is small.

Comparison with the Standard Branch-Site LRT

Bayesian model selection does not yield \( p \) values and does not require a formal decision rule to classify support for a model as significant or not significant. However, the use of formal decision rules allows us to refer to the FPR and TPR of Bayesian tests and allows comparison with the branch-site LRT under the \( p < 0.05 \) decision rule. In this article, we examined the FPR and TPR of the Bayesian version of the branch-site test using the criteria \( BF > 20:1 \) and \( BF > 3:1 \). Note that unlike classical significance testing, these criteria allow the possibility that the researcher will accept \( H_0 \) accept \( H_1 \), or neither. For example, if \( BF_{10} < 1:20 \) then \( H_0 \) will be accepted under either decision rule, which cannot occur under the LRT.

The Bayesian and branch-site LRTs have similar trade-offs between FPR and TPR as illustrated by their ROC curves. However, the Bayesian criteria of \( BF > 3:1 \) and \( BF > 20:1 \) do not select the same points on these curves as the LRT criterion of \( p < 0.05 \). We explain this by noting that the \( p < 0.05 \) criterion is designed to limit the FPR, and low FPR is not the same as strong evidence in favor of positive selection. In fact, the significance threshold of \( p < 0.05 \) frequently corresponds to evidence thresholds between 3:1 and 5:1 in favor of \( H_1 \) and to PPs of \( H_0 \) between 0.16 and 0.25 (Sellke et al. 2001). Such high probabilities that \( H_0 \) is true even when it has been rejected have been invoked to explain the frequent failure of replication for scientific results when \( H_0 \) is rejected with \( p \) values very close to 0.05 (Johnson 2013). Focusing on the FDR instead of the FPR may lead to more reliable conclusions. Further, focus on the FDR may allow easier comparison of Bayesian and frequentist tests, because PPs are actually similar to the FDR instead of to \( p \) values (Storey 2003).

We recommend that researchers use the more stringent \( BF > 20:1 \) criterion over the relatively weak \( BF > 3:1 \) criterion. We imagine that researchers could be hesitant to use the \( BF > 20:1 \) criterion because it may yield fewer significant tests than the \( BF > 3:1 \) and \( p < 0.05 \) criteria. However, our results indicate that although the \( BF > 20:1 \) criterion detects few genes containing positive selection where the evidence
for positive selection is weak, it detects a comparable number of genes to the branch-site LRT where the evidence for positive selection is stronger, as on the BS2 data set. Use of the BF > 3:1 and p < 0.05 criteria, on the other hand, may lead to large FDRs when the evidence for positive selection is weak. For example, on the BS1 data set, the branch-site LRT and the BF > 3:1 criteria both experience an FDR_{1:1} of greater than 30% despite having low FPRs. In contrast, the BF > 2:1 criterion detects no genes as containing positive selection, presumably because the evidence for positive selection is too weak.

Wider Implications

Although this study focuses on the branch-site model (Zhang et al. 2005), all methods that estimate positive selection from an excess of nonsynonymous substitutions would seem to be vulnerable to alignment errors. Incorporating tests such as that of Pond et al. (2011) into the joint estimation framework would be a natural next step. The framework presented in this article is not limited to positive selection but can be applied to any single-site property with an evolutionary model that specifies substitution rates between letters or codons. Future discoveries may enable multisite properties such as conserved DNA binding motifs to be incorporated into the statistical and evolutionary framework presented here.

More broadly, incorrectly aligning nonhomologous letters or codons may create spurious substitutions, leading to an elevated FPR and TPR for any site properties characterized by excess substitutions. On the other hand, site properties characterized by conservation will have a decreased FPR and TPR in the presence of alignment error if conserved columns are not correctly assembled. In such cases, we predict that integrating out the alignment will improve power by increasing a low TPR instead of by decreasing a high FPR. Censoring of misaligned regions seems unlikely to improve the ability to detect conserved sites, such as DNA binding motifs, when the conserved sites are themselves misaligned.

In view of the high accuracy and practical run time for alignment integration, we recommend that researchers who seek to infer site properties from sequence data should consider not only procedures for annotating and censoring alignments but also methods for integrating over them.

Materials and Methods

Model

Our model of the evolutionary process can be described in terms of the probability expression for the observed data and other unobserved components of the evolutionary process. The observed data consists of n observed sequences Yi, for i = 1...n. The phylogeny relating these sequences has unrooted topology τ and branch lengths T. Each observed sequence Yi corresponds to a leaf of the topology τ. The alignment expresses the positional homology of residues in these n observed leaf sequences and also the n-2 unobserved sequences at internal nodes. Evolutionary parameters Θ and Λ describe the process of accumulation of substitution and insertion/deletion mutations, respectively. Given this notation, we can describe the joint probability of all these random variables as:

\[
Pr(Y,A,\tau,T,\Theta,\Lambda) = Pr(Y | A,\tau,T,\Theta) \times Pr(A | \tau,T,\Lambda) \times Pr(\tau,T) \times Pr(\Theta) \times Pr(\Lambda).
\]

Here, the term \(Pr(Y | A,\tau,T,\Theta)\) is the standard substitution likelihood and is given by the substitution model. The term \(Pr(A | \tau,T,\Lambda)\) is given by the insertion–deletion model. The remaining terms are prior distributions on the phylogeny and evolutionary process parameters. In this model, the substitution process and insertion–deletion process operate completely independently from each other. This means that the rates of insertion and deletions are not influenced by what letters are inserted or deleted and that the rates of substitution are not affected by the presence of insertions or deletions.

Substitution Model

We make use of the branch-site model introduced by Zhang et al. (2005). As described above, model parameters include the frequencies \(p_0, p_1, p_2a,\) and \(p_2b\) of each site class, the frequencies of the 61 sense codons, the transition/transversion rate ratio \(\kappa\), and the nonsynonymous/synonymous rate ratios. We choose to parameterize the site class frequencies in terms of the relative frequencies \(f_1 = \frac{p_0}{p_0 + p_1}\) and \(f_2 = \frac{p_1}{p_0 + p_1}\) of each conserved or neutral site class, along with the fraction \(f_+ = \frac{p_2a + p_2b}{p_2a + p_2b}\) of positively selected sites. Thus, \(p_2a = f_+ \cdot f_1\) and \(p_2b = f_+ \cdot f_2\). Corresponding to these frequency parameters, we write \(\omega_0,\omega_1 = 1\), and \(\omega_+\), for \(\omega_0,\omega_1 = 1\) and \(\omega_+\), of Zhang et al. (2005). We make use of the F3x4 parameterization of codon frequencies. This parameterization determines the codon frequencies from independent nucleotide frequencies \(\pi^{(1)}, \pi^{(2)},\) and \(\pi^{(3)}\) in each codon position, renormalized to sum to 1.0 after the removal of the three stop codons.

We also introduce a binary indicator variable H to select between the null model with no positive selection and the alternative model with positive selection. When \(H = 0\), we ignore the value of \(\omega_+\) parameter and compute transition matrices as if \(\omega_+ = 1\). This corresponds to a lack of positive selection, although it still imposes a rate change from conservation to neutrality on foreground branches in site class #3 (table 2). When \(H = 1\), the value of the \(\omega_+\) parameter is used when computing transition matrices. Because this value is always greater than 1.0, this ensures a rate change to positive selection on the foreground branch in site classes #3 and #4.

The substitution model parameters are \(\Theta = (f_1, f_2, f_+, \omega_0, \omega_+ , \pi^{(1)}, \pi^{(2)}, \pi^{(3)}, \kappa, H)\), for a total number of 14 degrees of freedom.

Insertion–Deletion Model

We make use of the Redelings and Suchard (2007, RS07) model of insertion and deletion. This model constructs a distribution on multiple alignments from a collection of
pairwise alignment distributions placed along the branches of a phylogenetic tree. The pairwise alignment distributions are described by a pair-hidden Markov model. These pairwise alignment distributions are symmetrical in the ancestor and descendant sequences. This means that the indel model is reversible and that insertions and deletions are equally probable. The RS07 model allows multiresidue indels and thus has an affine gap penalty. Insertion and deletion lengths follow a common geometric length distribution with extension probability $\varepsilon$, so that the mean indel length is $1/(1-\varepsilon)$. Indels in the RS07 model occur at a rate $\lambda$, scaled relative to the substitution rate. Thus, the insertion–deletion parameters $\Lambda = (\lambda, \varepsilon)$ contribute 2 degrees of freedom.

Simulations

We simulated 1,000 data sets from each of six simulation conditions. The FY, BS1, and BS2—simulation conditions do not contain positive selection, whereas the FY+, BS1+, and BS2+ simulations contain positive selection at some fraction of sites on a single branch. All simulation regimes make use of a common rooted tree (fig. 5). The two branches connecting to the root are foreground branches, and the remaining branches are background branches. Simulations on the tree began with a sequence of 300 codons at the root. The transition/transversion rate ratio $\kappa$ was set to four on all branches. Codon frequencies were assigned based on the F3x4 model; nucleotide frequencies for the 1st, 2nd, and 3rd position were set to the same values used by Fletcher and Yang (2010). The insertion rate and the deletion rate were both set to 0.05 times the substitution rate. The length of both insertions and deletions followed a geometric distribution with success probability 0.35, so that the average indel length was 1.53 codons. All data sets were simulated using the software INDELible Fletcher and Yang (2009).

The FY simulation conditions are taken from Fletcher and Yang (2010). In the FY+ simulation conditions, each codon position has probability 1/10 of being assigned to each of 10 site classes. Each site class is assigned an $\omega$ value for background branches and an $\omega$ value for foreground branches according to table 3. These $\omega$ values correspond to schemes X and U from Fletcher and Yang (2010). On the background branches, each $\omega$ value 0.0, 0.2, 0.5, 0.8, or 1.0 is assigned to two site classes, and so these $\omega$ values each occur in 20% of alignment columns. On the foreground branches the only difference is that one of the two $\omega = 0.5$ site classes is changed to have $\omega = 2.0$, and one of the two $\omega = 0.8$ site classes is changed to have $\omega = 4.0$. Thus, 20% of sites switch to $\omega > 1$ on the foreground branch. However, columns with the highest conservation on background branches never switch to positive selection under the FY+ simulation conditions. This violates the assumptions of the branch-site model. The FY− model is derived from the FY+ model by making all branches into background branches. This differs from the null model of the branch-site test, which retains foreground branches but sets $\omega = 1.0$.

Because the FY simulation conditions violate the assumptions of the branch-site model, we introduce simulation conditions BS1+, BS1−, BS2+, and BS2−—that can be expressed in terms of the branch-site model. The BS data sets use the same tree, insertion–deletion parameters, and codon frequencies as the FY data sets. For all BS data sets, $f_+ = 0.2$, so that 20% of sites switch to positive selection on the foreground branch. For the BS1+ data set, we set $f_1 = 0.5$, $\omega_1 = 0.5, f_2 = 0.5, \omega_2 = 1.0$. For the BS2+ data set, we set $f_1 = 0.6, \omega_1 = 0.1, f_2 = 0.4, \omega_2 = 1.0$. The BS1− and BS2− models are derived from the BS1+ and BS2+ models by setting $\omega_+ = 1.0$. Thus, in the BS1− and BS2− models, rate switching does occur on the foreground branch. However, instead of switching to positive selection, the BS1− and BS2− allow only switching to neutrality.

Alignment Methods

We performed the Bayesian version of the branch-site test on each simulated data set using a variety of different alignment methods. Several methods relied on fixed, externally supplied alignments. These include the known true alignment, as well as alignments constructed by the software packages MUSCLE, MAFFT, FSA, and PRANK. Additionally, we refer to the results of the analysis in which the presence of positive selection and the alignment were jointly estimated as “Joint A.” An additional method involved selecting the last sampled alignment from the Joint A analysis and using it as input to a fixed-alignment analysis. We refer to this fixed-alignment analysis as “Joint F,” because only a single alignment was used. The exact commands are provided in the supplementary material, Supplementary Material online.

For the software packages MUSCLE, MAFFT, and FSA, data sets were aligned on the amino acid level to obtain alignments that do not split codons (Fletcher and Yang 2010). However, the PRANK software also contains the ability to align codons directly, and we therefore use codon-based alignments instead of amino-acid-based alignments from PRANK unless otherwise specified.

Priors

The Bayesian approach requires the incorporation of prior distributions for each parameter. As mentioned above, the priors on $\omega_+$ and $f_+$ are probably the most influential. We therefore construct priors on $\omega_+$ and $f_+$ that are sufficiently vague that they can be reused in future analyses of other data sets with different parameter values.

We place a $\Gamma(4,0.25)$ prior distribution on log $\omega_+$. We chose this prior to satisfy three important criteria. First, the distribution is vague and has a heavy right tail. This means that a broad range of $\omega_+$ values is plausible a priori. The heavy right tail means that the prior belief against large $\omega_+$ values is weak enough that large $\omega_+$ values can be inferred if the data support them. Second, the prior places about 50% of its mass between biologically plausible values between $\omega_+ = 2$ and $\omega_+ = 4$. Third, the prior density decreases to 0 as it approaches 1.0. This means the test will require more data to infer positive selection when $\omega_+$ is only slightly larger.
than 1. Finally, note that any prior on $\omega_+$ must have $\Pr(\omega_+ > 1) = 1$.

We place a $\beta(1,10)$ prior distribution on $f_+$. This distribution was chosen to satisfy three criteria. First, the prior mean of $f_+$ should not be too close to 0. If the prior mean of $f_+$ is too small, then it is possible to infer that a very small number of sites have a very high value of $\omega_+$. Second, the prior mean of $f_+$ should not be too large. If $f_+$ is estimated as being much larger than the true value, then neutral and conserved sites will be classified as positively selected. This will push the value of $\omega_+$ down and power will be lost. Finally, we also sought a prior that represents relatively weak evidence against large $f_+$ values, so that if the data set actually contains a high frequency of positively selected sites, estimation of a high value of $f_+$ will be possible.

We place a uniform prior on $H$, so that $\Pr(H = 0)$ and $\Pr(H = 1)$ are both 0.5. Because the likelihood is calculated as if $\omega_+ = 1$ when $H = 0$, this leads to a prior on $\omega_+$ that consists of 50% of the mass being placed on $\omega_+ = 1$ and 50% of the mass being placed on $\omega_+ > 1$ (fig. 8).

We place a Dirichlet $(1, 1)$ distribution on $(f_{\sigma 1})$. We place a Laplace($-4, 1/\sqrt{2}$) prior on the log of the insertion/deletion rate $\lambda$. We place an Exponential prior with mean 10 on the mean indel length minus 1. Because the topology is fixed, we do not need to place a prior on topologies. However, branch lengths are random and so we place a hierarchical prior on branch lengths, with each branch length $T_b \sim \Gamma(1/2, 2\mu)$ and the hyper parameter $\mu \sim \Gamma(1/2, 2)$. Thus, each branch length has prior mean $\mu$, and $\mu$ has prior mean 1.0. This hierarchical prior avoids sensitivity to the prior mean on branch lengths.

**Transition Kernels**

For continuous variables, we made use of both Metropolis–Hastings transition kernels and autotuned slice-sampling transition kernels. For the binary variable $H$, we used a simple Metropolis transition kernel to propose the alternative state. For the alignment, we made use of four main transition kernels. These include the HB1 and HB2 transition kernels (Holmes and Bruno 2001), along with two transition kernels described by Redelings and Suchard (2005). Each of these transition kernels resamples part of the alignment but keeps the remainder unchanged.

**MCMC Convergence**

Posterior samples were obtained by using the software **BAli-Phy** to perform MCMC. When estimating the alignment, the initial alignment was obtained by removing all gaps from the FASTA file containing the true alignment, thus resulting in an alignment with no internal gaps and external gaps only on the right edge. We ran two independent chains for each analysis and pooled the results. Each chain was run for 2,000 iterations, discarding the first 500 iterations as burnin. Samples are recorded once every iteration. Note that **BAli-Phy** performs a large number of operations in each iteration, so that **BAli-Phy** iterations are not necessarily comparable to iterations of other MCMC software. For example, every parameter and branch length was resampled in each iteration, and the pairwise alignment along each branch was resampled five times each iteration. Each chain requires about 15 h for 2,000 iterations on an Intel Xeon 5550 processor. This can be compared with a total time of about 10 min for **PRANK + GUIDANCE + CodeML**.

Convergence and mixing were assessed by examining the potential scale reduction factors (PSRF) based on the length of 80% credible intervals (Brooks and Gelman 1998). We examined the PSRF for all continuous parameters. In analyses where the alignment was estimated, we also examined the PSRF for the total number of indels, the total lengths of indels, the total number of alignment columns, and the nucleotide-wise parsimony score (Gaya et al. 2010). The median of PSRF across MCMC runs with a fixed alignment was 1.02. For MCMC runs where the alignment was being estimated, the median PSRF was 1.04.

We also measured the correlation between PPs of positive selection estimated from different MCMC runs. This correlation was 0.993 when the alignment was fixed to the true alignment. When integrating out the alignment and sampling the alignment only once per iteration, the correlation was 0.991. When increasing the alignment sampling by a factor of 5, as in the final results, the correlation increased to 0.992.

**Alignment Distances**

For a pair of sequences $i$ and $j$, the distance between two pairwise alignments $\alpha_1$ and $\alpha_2$ is computed as follows. We refer to letters of $i$ and $j$ by their position in the sequence, not by their value. Thus, for example, in the nucleotide sequence ATGA, the two As are considered different letters because they occur at different positions. Then let $d_1(\alpha_1, \alpha_2)$ be the number of letters in $i$ that are aligned differently between $\alpha_1$ and $\alpha_2$. This includes letters in $i$ that are aligned to a gap in one alignment but not the other, as well as letters in $i$ that are aligned to two different letters of $j$ in the two alignments. Similarly, let $d_2(\alpha_1, \alpha_2)$ be the number of letters of $j$ that are aligned differently between the two alignments. Furthermore, let $|i|$ and $|j|$ be the number of letters in

![Fig. 8. Prior distribution on $\omega_+$. The prior places 50% of its mass on 1. For the other 50% of the mass, the prior mean is about 3. The prior places low support on values >1.0 that are very close to 1.0. The prior has a heavy right tail, indicating that it does not strongly conflict with values of $\omega_+$ that are larger than the mean.](image-url)
the sequences $i$ and $j$, respectively. Then our distance $d(\alpha_i, \alpha_j)$ is defined to be:

$$d(\alpha_i, \alpha_j) = \frac{d_1(\alpha_i, \alpha_j) + d_2(\alpha_i, \alpha_j)}{|i| + |j|}.$$  

This distance is symmetric in $i$ and $j$, as well as symmetric in $\alpha_1$ and $\alpha_2$. Its value must be in the interval $[0,1]$. Unlike some other distances for pairwise alignments, this distance function rewards correct gaps and penalizes incorrect matches, in addition to rewarding correct matches (Bradley et al. 2009).

Software

All Bayesian analyses in this article were performed using the software BAli–Phy. Source code is freely available at https://github.com/bredelings/BAli-Phy (last accessed June 9, 2014).

Supplementary Material

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by National Science Foundation Grant #EF-0905606 to the National Evolutionary Synthesis Center (NESCent) and by Public Health Service grant #GM-37841. The author thanks Bill Fletcher for providing the INDELible control files used by Fletcher and Yang (2010). Many thanks to Ziheng Yang for his invaluable assistance in describing implementation details of CodeML. He also thanks three anonymous reviewers for their valuable assistance in improving the manuscript.

References


