Erasing Errors Due to Alignment Ambiguity When Estimating Positive Selection

Authors
Benjamin Redelings[1,2]
[1]Biology Department, Duke University
[2]National Evolutionary Synthesis Center (NESCent)

Corresponding author: Benjamin Redelings (*benjamin.redelings@duke.edu*)

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 MCMC 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 `BAli-Phy` have substantially lower alignment error compared to `MUSCLE`, `MAFFT`, `PRANK`, and `FSA` alignments.

Keywords: 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 (Muse and Gaut, 1994; Goldman and Yang, 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, since aligning non-homologous residues is likely to imply a spurious non-synonymous substitution, which will be interpreted as evidence for positive selection. Alignment errors, together with sequencing errors and the inclusion of non-homologous 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 in inferring positive selection (Fletcher and Yang, 2010), even in methods that have quite conservative false-positive rates when the true alignment is known.

In order 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` (Edgar, 2004) and `MAFFT` (Katoh et al., 2002; Katoh and Standley, 2013), both of which were superior to `ClustalW`. Researchers have also developed a wide variety of methods for detecting and removing unreliable regions from alignment estimates in order to decrease downstream false-positive rates. For example, `GBLOCKS` censors columns that are highly variable or near a gap (Castresana, 2000). `SOAP` determines reliability based on sensitivity to gap cost parameters (Löytynoja and Milinkovitch, 2001). `ALISCORE` compares the best alignment of letters within a window to the best alignment when those letters are randomly reordered (Misof and Misof, 2009). `GUIDANCE` (Penn et al., 2010a,b) measures sensitivity to the guide tree used in progressive alignment. `HoT` looks for differences between co-optimal alignments (Landan and Graur, 2008). Censoring methods such as these are able to improve the accuracy of site-wise detection of positive selection for less accurate alignment methods, but have a much smaller effect on more accurate methods such as `PRANK` (Jordan and Goldman, 2012; Privman et al., 2012).

More recently, researchers have adjusted likelihood ratio tests for positive selection by replacing likelihoods based on a single alignment with a likelihood averaged across a number of alignments taken from MCMC software such as `BAli-Phy` (Blackburne and Whelan, 2013).

Tab. 1: The M2a model for site-dependent ω.

	Class 1	Class 2	Class3
ω	$0 \leq \omega_0 \leq 1$	$\omega_1 = 1$	$\omega_2 \geq 1$
Frequency	p_0	p_1	$1 - p_0 - p_1$

These results suggest that a single posterior sample from `BAli-Phy` under the M0/RS07 model leads to nearly the same true positive and false positive rates as a single alignment from `PRANK`. However, the use of averaged likelihoods leads to a slight but measurable improvement in both the true positive and false positive rates.

Diversifying positive selection and the branch-site model

Diversifying positive selection is a property of codons, not of individual nucleotides. We therefore choose to focus on codon sites instead of nucleotide sites. The simplest way to explain diversifying positive selection is to write down the expression for the rate of substitution from one codon state to another, following the Goldman and Yang (1994, GY94) model. The GY94 model requires that codons may only change one nucleotide at a time. Subject to that constraint, the rate of substitution from one codon i to another codon j is given as

$$Q_{i\to j} = \pi_j \times \left\{ \begin{array}{cc} 1 & \text{if } \textit{transversion} \\ \kappa & \text{if } \textit{transition} \end{array} \right\} \times \left\{ \begin{array}{cc} 1 & \text{if } \textit{synonymous} \\ \omega & \text{if } \textit{non-synonymous} \end{array} \right\},$$

where π_j is the equilibrium frequency of codon j. Thus, the non-synonymous/synonymous (dN/dS) rate ratio ω represents an increased or decreased rate of change for nucleotide substitutions that result in amino acid changes, relative to what would be expected for neutral evolution. Thus if $\omega = 1$, we describe the process as neutral. If $\omega < 1$ we say that the codon is conserved and is undergoing negative selection. 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*.

In order 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 ω values. These ω 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 $\omega = \omega_2 \geq 1$. One can obtain a model without positive selection by constraining $\omega_2 = 1$ and $1-p_0-p_1 = 0$, and this leads to a likelihood ratio test (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 were tested separately. (Wong et al., 2004). However, note that if we mistakenly align two separate sub-columns into a single (incorrect) column, then we may create a spurious non-synonymous substitution. Since 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.

Tab. 2: The branch-site model for branch- and site-dependent ω.

	Class 1	Class 2	Class 3	Class 4
background	ω_0	1	ω_0	1
foreground	ω_0	1	ω_2	ω_2
Frequency	p_0	p_1	p_{2a}	p_{2b}

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, while the remainder are labeled "background" branches. In this "branch-site" model, the ω for a site may switch to $\omega_2 \geq 1$ on the foreground branches, which remain either $\omega_0 \leq 1$ or $\omega_1 = 1$ on all background branches. Some fraction of sites p_2 undergo this switch, and that fraction is chosen independently of the ω value for the background branches. Zhang et al. (2005) suggest constructing a LRT by comparing this model with a null model where ω_2 is constrained to 1. This null model is preferred over the null model where $1 - p_0 - p_1$ is constrained to be 0, since it allows ω 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 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 false positive rates for the branch-site test that are substantially higher than 0.05. Following Zhang et al. (2005) and Zhang (2004), Fletcher and Yang used a simulation model that allowed columns to select from a variety of different neutral or conservative ω values on background branches. It also allowed positive selection on the foreground branch to be strongly correlated with the ω value on background branches. Fletcher and Yang also systematically varied the location of the foreground branch, the tree shape, and the distribution of ω values on the foreground branch.

Despite the fact that these simulation models contra-

Tab. 3: Site categories and their ω values under different selection schemes.

	Scheme	
Site Class	X	U
1	1.0	1.0
2	1.0	1.0
3	0.8	4.0
4	0.8	0.8
5	0.5	2.0
6	0.5	0.5
7	0.2	0.2
8	0.2	0.2
9	0.0	0.0
10	0.0	0.0

dict 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 false positive rates 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 false positive rates as high as 0.99. Other alignment software performed better, with `PRANK` codon-based alignments having the lowest false positive rates. Nevertheless, `PRANK` alignments had false positive rates as high as 0.13 under some evolutionary scenarios, substantially exceeding the desired level of 0.05.

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 $\hat{\mathbf{A}}$ is not observed, and so is not considered data. This affects the likelihood, since the likelihood is defined to be proportional to the probability of the data, given hypothesis H and parameters Θ. The likelihood does not condition on $\hat{\mathbf{A}}$, since $\hat{\mathbf{A}}$ is not a model parameter.

$$\Pr(\text{unaligned data}|H,\Theta) \neq \Pr(\text{unaligned data}|H,\Theta,\hat{\mathbf{A}}).$$

Instead, the likelihood integrates over the alignment $\mathbf{A}$, since $\mathbf{A}$ is a latent variable:

$$\Pr(\text{unaligned data}|H,\Theta) = \sum_A \Pr(\text{unaligned data},\mathbf{A}|H,\Theta).$$

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

$$BF_{10} \quad = \quad \frac{\sum_A \Pr(\text{unaligned data},\mathbf{A}|H=1)}{\sum_A \Pr(\text{unaligned data},\mathbf{A}|H=0)}.$$

Here $H = 0$ indicates that the null model (H_0, no positive selection) is true, while $H = 1$ indicates that the alternative model (H_1, with positive selection) is true. In order 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 that $H = 1$, which we can use to compute the Bayes factor.

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 $\hat{\mathbf{A}}$. 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 in order 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, while 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_{10}, which is the Bayes factor in favor of H_1 against H_0, and thus quantifies the evidence in favor of positive selection.

In order 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_2, and the prior distribution on ω_2. These priors play a crucial role in defining H_1 because they determine what fraction p_2 of sites display positive selection under H_1, as well as the strength ω_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.

While Bayes factors can be computed without specifying the prior probability that $H = 0$ and $H = 1$, posterior probabilities (PP) 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 $\text{FDR}_{1:1}$. In this "equipoise" scenario, the Bayes factor equals the posterior odds. Posterior probabilities > 0.952 then correspond to a Bayes factor >20:1, whereas posterior probabilities >0.75 correspond to a Bayes factor >3:1.

In the rest of this paper, we first describe how to estimate the Bayes factor with sufficient accuracy. We then simulate data according to a 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 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 $\Pr(H_1|\text{data}) = \Pr(H = 1|\text{data})$, and the probability of no positive selection is $\Pr(H_0|\text{data}) = \Pr(H = 0|\text{data})$.

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

$$\Pr(H = 1|data) \approx \frac{1}{N}\sum_{j=1}^{N} 1_{\{h_j=1\}}. \qquad (1)$$

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 $\Pr(H = 1|data)$ 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 posterior probabilities may seem similar, they lead to the very different odds ratios of 99/1 and $100/0 = \infty$. 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 in order to compute high posterior odds without obtaining an enormous number of samples from the Markov chain.

Our strategy for obtaining improved estimates of $\Pr(H = 1|\text{data})$ is to record at each iteration not just 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 $\Pr(H = 1|\text{data})$, we define X to refer to all variables in the Markov chain except H. We note that

$$\begin{aligned}\Pr(H = 1|\text{data}) &= \text{E}\left(1_{\{H=1\}}\middle|\text{data}\right)\\ &= \text{E}\left[\text{E}\left(1_{\{H=1\}}\middle|X\right)\middle|\text{data}\right]\end{aligned}$$

$$= \quad \mathrm{E}\left[\Pr\left(H=1\middle|X\right)\middle|\mathrm{data}\right],$$

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 jth iteration of the Markov chain. Then the approximation

$$\Pr(H=1|\mathrm{data}) = \mathrm{E}\left[\Pr\left(H=1\middle|X\right)\middle|\mathrm{data}\right]$$

$$\approx \frac{1}{N}\sum_{j=1}^{N}\Pr\left(H=1\middle|X=x_j,\mathrm{data}\right) \qquad (2)$$

allows us to approximate $\Pr(H=1|\mathrm{data})$ by averaging over the value of $\Pr\left(H=1\middle|X=x_j,\mathrm{data}\right)$ that is recorded at each iteration.

In order to compute $\Pr\left(H=1\middle|X=x_j,\mathrm{data}\right)$, we modify the software to compute $\Pr(H=0,X=x_j,\mathrm{data})$ and $\Pr(H=1,X=x_j,\mathrm{data})$ at each iteration without changing h_j. Then

$$\Pr\left(H=1\middle|X=x_j,\mathrm{data}\right) = \frac{\Pr(H=1,X=x_j,\mathrm{data})}{\Pr(X=x_j,\mathrm{data})}$$

$$= \frac{\Pr(H=1,X=x_j,\mathrm{data})}{\Pr(H=0,X=x_j,\mathrm{data})+\Pr(H=1,X=x_j,\mathrm{data})}.$$

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 1000 data sets with diversifying positive selection throughout the entire gene region (FY+) and 1000 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 co-estimation analysis, PPs are nearly as accurate as when performing a full co-estimation 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, posterior probabilities are shifted toward 1.0. However, PPs under joint estimation have nearly the same distribution as when the true alignment is known.

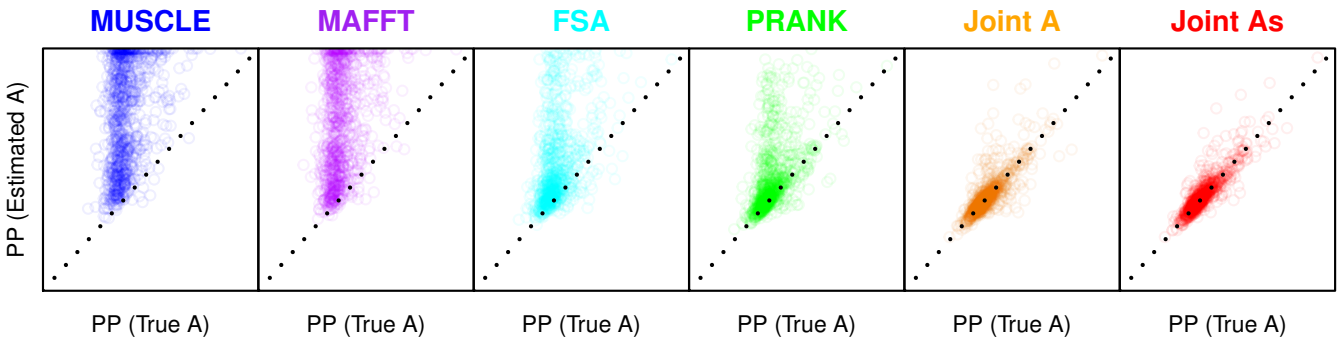
We calculate the squared correlation across data sets of PP for estimated alignments versus PP for true alignments. For data sets simulated under the FY- model without positive selection, the squared correlation coefficient is 0.088 for `MUSCLE`, 0.080 for `MAFFT`, 0.17 for `FSA`, 0.26 for `PRANK`, 0.66 when fixing a posterior sampled alignment, and 0.79 when co- estimating 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

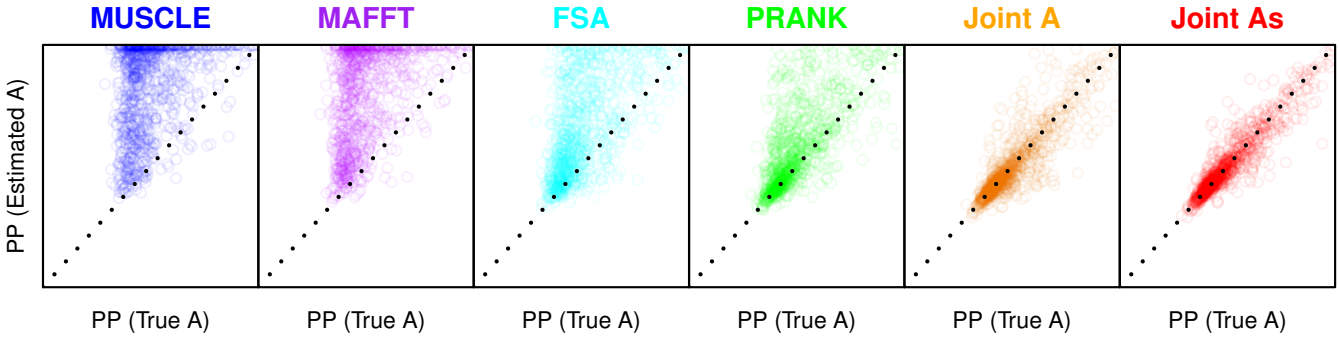
In order to assess the ability of different methods to discriminate between data sets with and without positive selection, we compute ROC curves for Bayesian inference of positive selection using different alignment methods (fig. 3). ROC curves allow comparison of different methods at the same level of FPR, even if those methods achieve different FPR values in practice. Depending on the method, the alignment was either co-estimated (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 co-estimation analysis. The TPR and FPR were computed based on the FY+ and FY- data sets, respectively (table 4). At a 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), while 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.

Joint estimation avoids inflated FPRs

Joint estimation eliminated excess FPRs for the Bayesian tests. However, the use of estimated alignments lead to inflated FPR values (table 4). For the 'True A', 'Joint As', and 'Joint A' analyses, Bayesian tests based on the



(a) No positive selection



(b) Positive selection

Fig. 1: The PP of positive selection given the true alignment (x-axis) versus the PP under various alignment estimation methods (y-axis). Plots are based on data sets simulated with positive selection (bottom row) or without positive selection (top row). Points falling above the black dotted line indicate excess confidence of positive selection because the alignment is not known *a priori*. PPs on the y-axis are based on alignments estimated using `MUSCLE`, `MAFFT`, `FSA`, `PRANK`, a sampled alignment (Joint A), or joint estimation averaging over alignments (Joint As) , as indicated.

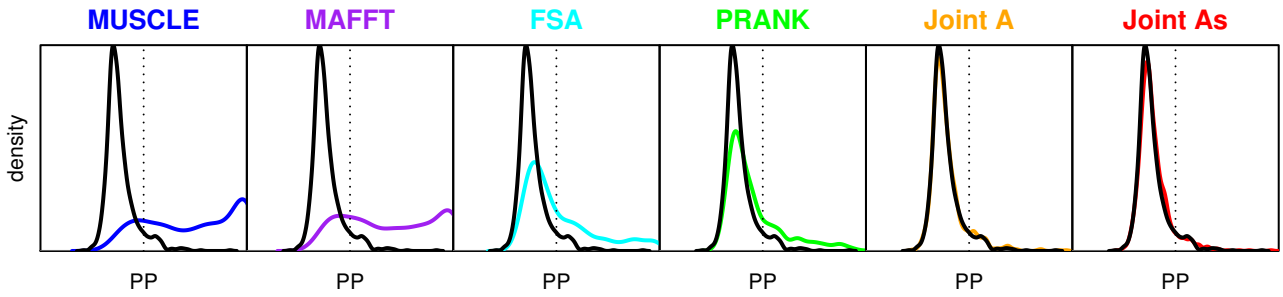
BF>3:1 and BF>20:1 criterion all yielded a FPR of <1%. Under the BF>3:1 criterion, estimating alignments with `PRANK`, `FSA`, `MAFFT`, or `MUSCLE` lead to FPRs of 6%, 13%, 43%, and 52% respectively. Under the BF>20:1 criterion, estimating alignments lead to FPRs of <1%, 3%, 21%, and 27% respectively.

Performance of the LRT and Bayesian branch-site tests

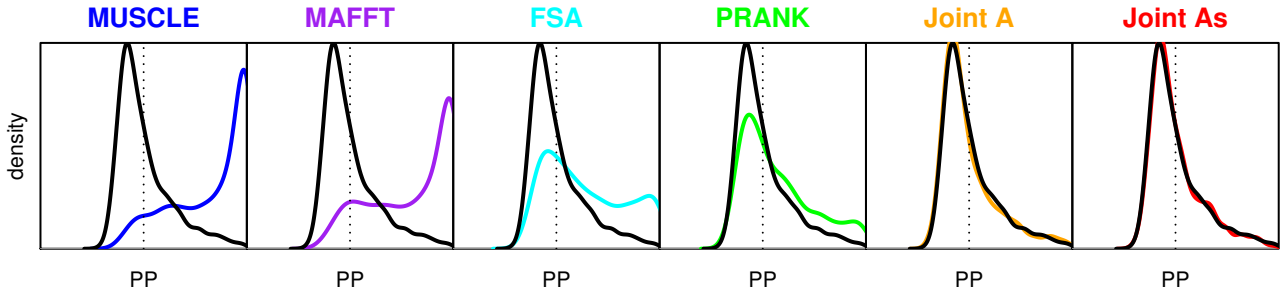
The Bayesian tests and the branch-site LRT have similar trade-offs between FPR and TPR on the FY, BS1, and BS2 simulated data sets as illustrated by their ROC curves (fig. 4). For comparisons between the Bayesian and LRT approaches we assume that the true alignment is known in order to focus on the difference in approach. The BS1 conditions yield little power to detect positive selection, the BS2 simulation conditions lead to higher power, and the FY simulation conditions are intermediate. The FY and BS1 data sets lead to nearly identical ROC curves for the Bayesian and LRT approaches. However, on the BS2 data set Bayesian inference leads to a ROC that clearly dominates the LRT curve. For example, at an FPR of 5%, the LRT has a TPR of 59% while 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 set, the standard branch-site LRT based on the conservative χ^2_1 distribution has a 1% FPR, a 13% TPR, and a 8% $\text{FDR}_{1:1}$. (For comparison, 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% $\text{FDR}_{1:1}$.) 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 $\text{FDR}_{1:1}$ of 3%, while 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.

Since 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% $\text{FDR}_{1:1}$. The BF>3:1 criterion for the Bayesian test attains an FPR <1%, a 3% TPR, and a 31%



(a) No positive selection



(b) Positive selection

Fig. 2: Distributions of the PP of positive selection across data sets simulated with positive selection (bottom row) or without positive selection (top row). The x-axis in each cell ranges from 0 to 1, while the y-axis indicates probability density. The solid black curve in each panel represents the distribution of PPs based on the true alignment. The other curve represents the distribution of PPs based on alignments estimated using `MUSCLE`, `MAFFT`, `FSA`, `PRANK`, a sampled alignment (Joint A) , or joint estimation averaging over alignments (Joint As), as indicated.

FDR, while 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 an 4% FPR, a 53% TPR, and 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% $FDR_{1:1}$.

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 7 can only be 0.2, 0.4, 0.6, or 0.8. Figure 5 plots the pairwise alignment error versus evolutionary distance for alignments estimated using `MUSCLE`, `MAFFT`, `FSA`, `PRANK` (dna), `PRANK` (aa), `PRANK` (codon), 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-data sets. Figure 6 shows the joint distribution of the true alignment length and estimated alignment length for `MUSCLE`, `MAFFT`, `FSA`, `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 re-

Tab. 4: Performance of Bayesian tests under different alignment estimates

		True A	Joint As	Joint A	PRANK	FSA	MAFFT	MUSCLE
FPR=5%	FPR	5%	5%	5%	5%	5%	5%	5%
	TPR	30%	27%	25%	15%	15%	11%	9%
	$FDR_{1:1}$	15%	16%	17%	25%	25%	31%	35%
FPR=1%	FPR	1%	1%	1%	1%	1%	1%	1%
	TPR	14%	13%	11%	7%	4%	3%	2%
	$FDR_{1:1}$	7%	9%	9%	12%	20%	29%	38%
BF>3:1	FPR	<1%	<1%	<1%	6%	13%	43%	52%
	TPR	7%	6%	7%	18%	31%	63%	70%
	$FDR_{1:1}$	3%	6%	7%	24%	29%	41%	42%
BF>20:1	FPR	<1%	<1%	<1%	<1%	3%	21%	27%
	TPR	<1%	<1%	<1%	4%	10%	38%	42%
	$FDR_{1:1}$	?	?	?	?	26%	36%	40%

spectively. Differences for PRANK alignments and posterior sampled alignments were +16, and +1 codon, respectively. Thus alignments sampled under the BST/RS07 model are nearly unbiased, while 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, while 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 false positive rates in estimating positive selection. Despite their similar performance in detecting positive selection, FSA and PRANK alignments have different characteristics, since FSA alignments tend to be shorter than the true alignment, while 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. In order to fully account for alignment uncertainty, a procedure must account for both near-optimal alignments and the effect of uncertainty 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-censoring methods which usually consider only one source of alignment uncertainty. For example, GUIDANCE (Penn et al., 2010a,b) 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 con-

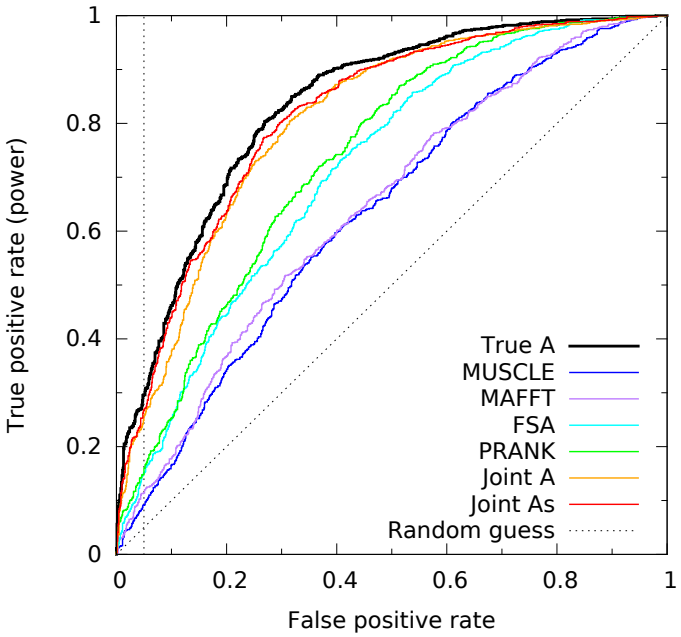


Fig. 3: ROC curves for inferring positive selection using different methods of alignment. Vertical dotted line indicates 5% FPR. Diagonal dotted line describes the performance of a random guess at different levels of specificity. Each curve represents Bayesian estimation based on alignments estimated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

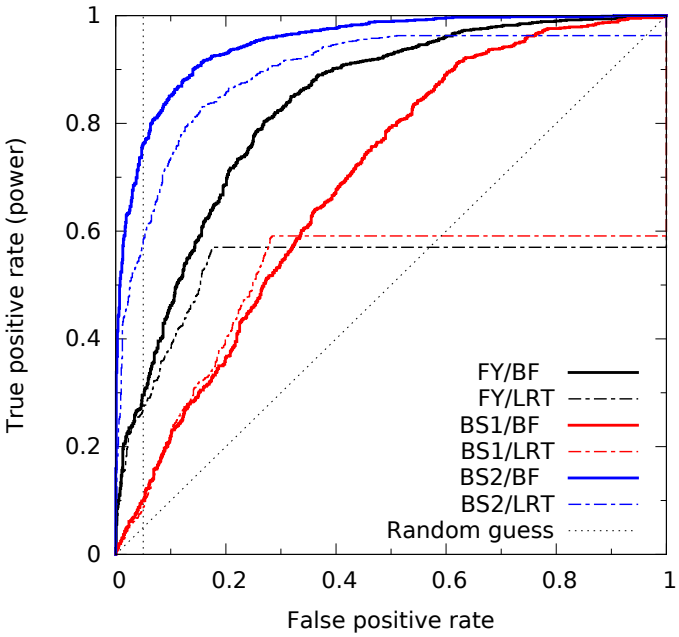


Fig. 4: ROC curves for Bayesian inference and for the branch-site LRT on the FY, BS1, and BS2 simulated data sets when the true alignment is known.

sider uncertainty resulting from uncertainty in parameters such as the phylogeny or gap penalties.

In this paper 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 `BAli-Phy` (Redelings and Suchard, 2005, 2007). However, the branch-site model prevents this, since 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 selecting to be co-estimated 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, which is already available in BAli-Phy.

Mismatches between models and reality

The currently study examines the effect of alignment error on inferring positive selection by focusing on a single set of simulation conditions (FY). Under these conditions, inferring positive selection based on a single sample from the posterior distribution was almost as accurate as the more rigorous method of performing an average over the sampled alignments. 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). Since we do not examine such conditions in the

Tab. 5: Performance of LRT and Bayesian tests on different simulated data sets

		FY	BS1	BS2
$p < 0.05$ (LRT)	FPR	1%	2%	4%
	TPR	13%	3%	53%
	$FDR_{1:1}$	8%	36%	6%
BF>3:1	FPR	<1%	<1%	9%
	TPR	7%	3%	84%
	$FDR_{1:1}$	3%	31%	10%
BF>20:1	FPR	<1%	<1%	<1%
	TPR	<1%	<1%	43%
	$FDR_{1:1}$	?	?	<1%

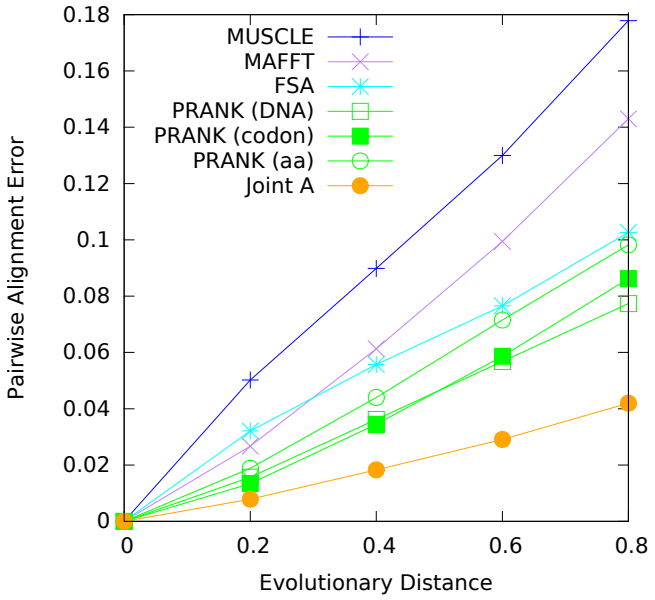


Fig. 5: Mean pairwise alignment error at various evolutionary distances. Pairs of leaf sequences with greater evolutionary distance have a greater degree of alignment error.

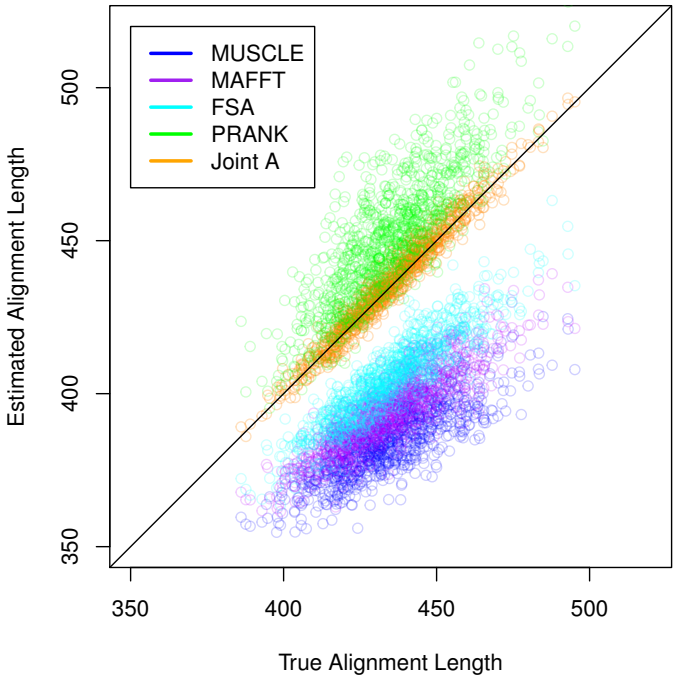


Fig. 6: Distribution of true and estimated alignment lengths for `MUSCLE`, `MAFFT`, `FSA`, `PRANK`, and samples from the BS/RS07 alignment posterior.

paper, 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 paper 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 paper. 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 π_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:

$$\begin{aligned} \pi_1 &\sim Uniform(0,1) \\ H_g &\sim Bernoulli(\pi_1) \text{ for } g = 1 \ldots G \end{aligned}$$

Such an approach would not be computationally prohibitive, since 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, since it would naturally require stronger evidence to infer positive selection when the fraction of genes experiencing positive selection is small. For data sets containing a large number of genes, it would also be possible to pool information about ω_+ and p_+ between genes using hierarchical priors to obtain more precise estimates.

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 paper, 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 tradeoffs 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 posterior probabilities 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, since posterior probabilities are actually similar to the false discovery rate (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 a $\text{FDR}_{1:1}$ of greater than 30% despite having low FPRs. In contrast, the BF>20:1 criterion detects no genes as containing positive selection, presumably because the evidence for positive selection is too weak.

Wider implications

While the current study focuses on the branch-site model (Zhang et al., 2005), all methods that estimate positive selection from an excess of non-synonymous 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 paper 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 multi-site properties such as conserved DNA binding motifs to be incorporated into the statistical and evolutionary framework presented here.

More broadly, incorrectly aligning non-homologous 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.

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 $\mathbf{Y}$ consists of n observed sequences $\mathbf{Y}_i$ for $i = 1 \ldots n$. The phylogeny relating these sequences has unrooted topology τ and branch lengths T. Each observed sequence $\mathbf{Y}_i$ corresponds to a leaf of the topology τ. The alignment $\mathbf{A}$ 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:

$$\Pr(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}, \Lambda) = \Pr(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta}) \times \Pr(\mathbf{A}|\tau, \mathbf{T}, \boldsymbol{\Lambda}) \times \Pr(\tau, \mathbf{T}) \times \Pr(\boldsymbol{\Theta}) \times \Pr(\boldsymbol{\Lambda}).$$

Here, the term $\Pr(\mathbf{Y}|\mathbf{A}, \tau, \mathbf{T}, \boldsymbol{\Theta})$ is the standard substitution likelihood, and is given by the substitution model. The term $\Pr(\mathbf{A}|\tau, \mathbf{T}, \boldsymbol{\Lambda})$ 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 κ, and the non-synonymous/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 selected site class, along with the fraction $f_+ = p_{2a} + p_{2b}$ 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 ω_1, $\omega_2 = 1$, and ω_+ for the ω_0, $\omega_1 = 1$ and ω_2 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 3 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 ω_+ 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 (see table 2). When $H = 1$, the value of the ω_+ parameter is used when computing transition matrices. Since 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 $\mathbf{\Theta} = (f_1, f_2, f_+, \omega_1, \omega_+, \pi^{(1)}, \pi^{(2)}, \pi^{(3)}, \kappa, H)$, 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 (pair-HMM). 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 multi-residue indels and thus has an affine gap penalty. Insertion and deletion lengths follow a common geometric length distribution with extension probability ϵ, so that the mean indel length is $1/(1-\epsilon)$. Indels in the RS07 model occur at a rate λ, scaled relative to the substitution rate. Thus the insertion-deletion parameters $\Lambda = (\lambda, \epsilon)$ contribute 2 degrees of freedom.

Simulations

We simulated datasets under models without positive selection and datasets under models containing positive selection on a single branch. For easy comparison with previous work, we used two simulation scenarios from Fletcher and Yang (2010). We refer to these scenarios as FY+ and FY- to indicate the presence or absence of positive selection. The FY data sets are not simulated under the branch-site model. Instead, they allow different columns to take on a variety of different neutral or conservative ω values on background branches. We therefore introduce data sets BS1+, BS1-, BS2+, and BS2- simulated under the branch-site model. We simulated 1000 data sets for each of the 6 simulation regimes. All data sets were simulated using the software `INDELible` Fletcher and Yang (2009).

All simulation regimes make use of a common rooted tree. The two branches connecting to the root are foreground branches, and the remaining branches are background branches. Simulations on the tree began with

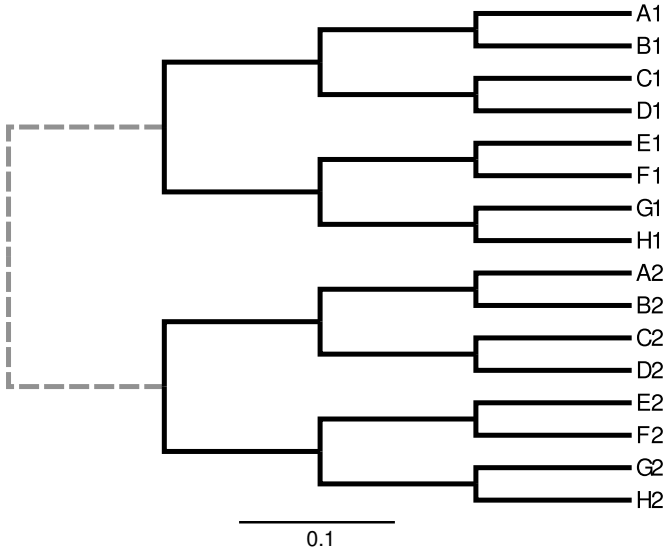


Fig. 7: Evolutionary tree used in simulation. The foreground branch is dashed and colored gray. It is referred to as branch α in Zhang et al. (2005). Branch lengths are given in terms of synonymous changes per synonymous site.

a sequence of 300 codons at the root. The transition/transversion rate ratio κ was set to 4 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.

In the FY data sets, each codon position had probability 1/10 of being assigned to each of 10 site classes. Each site class was assigned specific values of ω on each branch according to one of the two schemes X or U taken from Fletcher and Yang (2010). These selection schemes

are described in table 3. In the FY- model, all branches follow scheme X, whereas in the FY+ model the foreground branch is changed to follow scheme U. In scheme X, each ω value 0.0, 0.2, 0.5, 0.8, or 1.0 is assigned to 2 site classes, and so these ω values each occur in 20% of alignment columns. The U scheme differs from the X scheme only in 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, in the simulation scheme for positive selection, 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 in under the FY+ simulation conditions.

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 $f_1 = 0.5$, $\omega_1 = 0.5$, $f_2 = 0.5$, $\omega_2 = 1.0$, $\omega_+ = 3.0$. For the BS2+ data set, we set $f_1 = 0.6$, $\omega_1 = 0.1$, $f_2 = 0.4$, $\omega_2 = 1.0$, $\omega_+ = 4.0$. 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 As'. An additional method involved selecting the last sampled alignment from the 'Joint As' analysis and using it as input to a fixed-alignment analysis. We refer to this fixed-alignment analysis as 'Joint A', since only a single alignment was used.

For the software packages `MUSCLE`, `MAFFT`, and `FSA` data sets were aligned on the amino acid level in order 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 ω_+ and f_+ are probably the most influential. We therefore construct priors on ω_+ 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 on $\log \omega_+$ because this prior has three important properties. First, the distribution is vague and has a heavy right tail. This means that a broad range of ω_+ values is plausible *a priori*. The heavy right tail means that the prior belief against large ω_+ values is weak enough that large ω_+ 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 ω_+ is only slightly larger than 1. Finally, note that any prior on ω_+ must have $\Pr(\omega_+ > 1) = 1$.

We place a a Beta$(1, 10)$ distribution on p_+. We sought a prior that places most of its mass on values < 0.2. This is because if f_+ is estimated as being much larger than the true value, then the category of positively selected sites will effectively include neutral or conserved sites. This will then push the value of ω_+ down, and power will be lost. 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. Since the likelihood is calculated as if $\omega_+ = 1$ when $H = 0$, this leads to a prior on ω_+ that consists of 50% of the mass being placed on $\omega_+ = 1$, and 50% of the mass being placed on $\omega_+ > 1$ (fig. 8) .

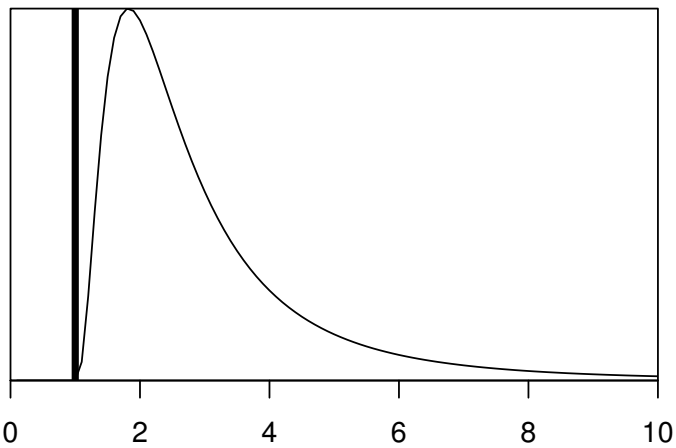


Fig. 8: Prior distribution on ω_+. The prior places 50% of its mass on 1. For the other 50% of the mass, the prior mean is about 3. The prior places low support on values > 1.0 that are very close to 1.0. The prior has a heavy right tail, indicating that it does not strongly conflict with values of ω_+ that are larger than the mean.

We place a Dirichlet$(1, 1)$ distribution on (f_1, f_2). We place a Laplace$(-4, 1/\sqrt{2})$ prior on the log of λ the rate of insertions and deletions. We place an Exponential prior

with mean 10 on the mean indel length minus 1. Since 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 μ, and μ 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 auto-tuned 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 4 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 re-samples 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 2 independent chains for each analysis and pooled the results. Each chain was run for 2000 iterations, discarding the first 500 iterations as burn-in. 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 re-sampled in each iteration, and the pairwise alignment along each branch was re-sampled 5 times each iteration. Each chain requires about 15 hours for 2000 iterations on an Intel Xeon 5550 processor. This can be compared to a total time of about 10 minutes 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 alignments 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 posterior probabilities 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 α_1 and α_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 sequence ATA, 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 α_1 and α_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. Likewise, 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 the sequences i and j respectively. Then our distance $d(\alpha_1, \alpha_2)$ is defined to be:

$$d(\alpha_1, \alpha_2) \quad = \quad \frac{d_1(\alpha_1, \alpha_2) + d_2(\alpha_1, \alpha_2)}{|i| + |j|}.$$

This distance is symmetric in i and j, as well as symmetric in α_1 and α_2. Its values 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`.

Acknowledgements

This work was supported by NSF Grant #EF-0905606 to the National Evolutionary Synthesis Center (NESCent). We thank 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`. I also thank the three anonymous reviewers for their help in improving the manuscript.

References

Allison L, Wallace CS. 1994. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and the opti-

misation of multiple alignments. *Journal of Molecular Evolution*. 39:418–430.

Blackburne BP, Whelan S. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol Biol Evol*. 30:642–653.

Blackwell D. 1947. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*. 18:1–164.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS computational biology*. 5:e1000392.

Brooks S, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*. 7:434–455.

Cartwright RA. 2006. Logarithmic gap costs decrease alignment accuracy. *BMC Bioinformatics*. 7:527.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *J. Mol. Biol. Evol.* 17:540–552.

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

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

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*. 27:2257–2267.

Gaya E, Redelings BD, Navarro-Rosiné P, Llimona X, Cáeres MD, Lutzoni FM. 2010. Align, or not to align? Resolving species complexes within the *Caloplaca saxicola* group as a case study. *Mycologia*. .

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

Golenberg EM, Clegg MT, Durbin ML, Doebly D, Ma DP. 1993. Evolution of a noncoding region of the chloroplast genome. *Molecular Phylogenetics and Evolution*. 2:52–64.

Holmes I, Bruno WJ. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*. 17:802–820.

Jeffreys H. 1998. The theory of probability. Oxford University Press.

Johnson VE. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*. 110:19313–19317.

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

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the american statistical association*. 90:773–795.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*. 30:3059–3066.

Katoh K, Standley DM. 2013. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.

Landan G, Graur D. 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac Symp Biocomput*. pp. 15–24.

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, Milinkovitch MC. 2001. Soap, cleaning multiple alignments from unstable blocks. *Bioinformatics*. 17:573–574.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome Res*. 21:863–874.

Misof B, Misof K. 2009. A monte carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst Biol*. 58:21–34.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*. 11:715–724.

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.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010a. Guidance: a web server for assessing alignment confidence scores. *Nucleic Acids Res*. 38:W23–W28.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010b. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*. 27:1759–1767.

Pond SLK, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. 28:3033–3043.

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

Redelings BD, Suchard MA. 2005. Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology.* 54:401–418.

Redelings BD, Suchard MA. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evo. Bio.* 7. 40.

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.* 1:114–118.

Sellke T, Bayarri M, Berger JO. 2001. Calibration of ρ values for testing precise null hypotheses. *The American Statistician.* 55:62–71.

Siepel A, Bejerano G, Pedersen JS, et al. (11 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research.* 15:1034–1050.

Sinha S, Blanchette M, Tompa M. 2004. Phyme: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC bioinformatics.* 5:170.

Storey JD. 2003. The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of statistics.* pp. 2013–2035.

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time markov chain evolutionary models. *Molecular Biology and Evolution.* 18:1001–1013.

Thorne JL, Kishino H. 1992. Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution.* 9:1148–1162.

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol.* 5:457–467.

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

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics.* 168:1041–1051.

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, 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.