(working title) Filtering with Guidance is pretty useless

Stephanie J. Spielman[1*] and Eric T. Dawson[1] Claus O. Wilke[1]

Address:
[1]Section of Integrative Biology and Center for Computational Biology and Bioinformatics. The University of Texas at Austin, Austin, TX 78731, USA.

[*]Corresponding author
Email: stephanie.spielman@utexas.edu

Manuscript type: research article

Keywords: multiple sequence alignment, alignment filters, sequence simulation, positive selection inference

Abstract

Most nucleotide or amino acid sequence studies begin with building a multiple sequence alignment (MSA) with the goal of assigning homology across positions of these diverged sequences. Although a critical step in nearly sequence evolution studies, MSAs may represent a substantial source of error in subsequent applications, ranging from phylogenetic reconstruction to evolutionary rate inference. Motivated by this issue, many have suggested methods which can identify putatively poorly aligned regions in MSAs so that researchers can remove these regions from analyses in an effort to reduce error. One recent method, known as Guidance, generates position-based confidence scores in a given MSA using a bootstrap approach. Some studies have suggested that applying a Guidance filter to protein-coding sequence alignments can improve accuracy in positive selection inference. However, the specific circumstances under which Guidance might offer improvements remain unclear. To elucidate how the Guidance filter behaves in realistic protein-coding sequences, we have re-implemented the Guidance software with several statistical improvements, including methods to phylogenetically correct confidence scores. We find that, overall, neither the original Guidance nor our phylogenetically-corrected version substantially improves positive selection inference; in certain cases, filtering can actually hinder accuracy. Thus, we cannot equivocally advocate the use of a Guidance-based filter for positive selection inference.

Introduction

Constructing a multiple sequence alignment (MSA) represents the first step of analysis in most studies of molecular evolution, namely phylogenetic reconstruction and evolutionary rate inference. Recently, several studies have shown that poor MSA quality can significantly hinder accuracy in such downstream analyses (Jordan and Goldman 2011; Markova-Raina and Petrov 2011; Dwivedi and Gadagkar 2009; Talavera and Castresana 2007; Ogden and Rosenberg 2006). In particular, using low-quality MSAs during positive selection inference tends to yield elevated false positive rates (Jordan and Goldman 2011; Privman et al. 2012; Schneider et al. 2009; Fletcher and Yang 2010). As a consequence, many have advocated applying an alignment filter to MSAs before their use in such analyses. Such filters, which include Guidance (Penn

et al. 2010; Privman et al. 2012), GBLOCKS (Castresana 2000) and T-Coffee (Notredame et al. 2000), locate putatively poorly aligned regions in alignments, thereby allowing users to curate their alignments to maximize signal. The hope is that culling unreliable positions and/or columns from MSAs will yield increased accuracy in positive selection inference, without excessively sacrificing power.

Of these filters, Guidance (Penn et al. 2010) is currently the most widely used in positive selection inference. Guidance derives a confidence score for each MSA position using a bootstrap approach that samples variants in a progressive alignment guide tree. Using the resulting confidence scores, users can "mask" (i.e. replace with an ambiguous character such as "?" or "N") positions which score below a given threshold, thereby removing residues that cannot be confidently placed in the alignment. This method is fairly conservative, as particular positions of low confidence can be removed rather than entire columns. Moreover, recent simulation studies have demonstrated that filtering poorly aligned residues with Guidance may confer increased accuracy when inferring positive selection (Jordan and Goldman 2011; Privman et al. 2012).

Even so, while Privman et al. (Privman et al. 2012) found dramatic improvements in positive selection inference when applying the Guidance filter, Jordan and Goldman's (Jordan and Goldman 2011) comprehensive study on alignment methods and filtering found that Guidance yields modest, if any, effects on this inference. In particular, they reported that, at typical protein divergence levels or insertion/deletion (indel) rates, Guidance offers few benefits to positive selection inference; while the Guidance filter did not obscure signal, neither did it improve inferences. Alternatively, at extremely high divergence levels (1.8 mean path length, defined as mean root-to-tip branch length) or indel rates (0.2 indel/substitution events), Jordan and Goldman found that Guidance significantly boosted true positive rates. While these results were compelling, protein sequences used in positive selection studies, however, rarely, if ever, contain sequences separated by such high divergences; for example, a typical mammalian gene tree's MPL is only about 0.17 (Spielman and Wilke 2013), 10% the divergence level at which Jordan and Goldman detected improvements when using the Guidance filter.

Motivated by these apparent discrepancies, we sought to determine whether altering the Guidance scoring algorithm could yield improved inferences at

more realistic divergence and/or indel levels. To this end, we re-implemented the Guidance software (see Methods for details) and examined the effects of new scoring algorithms which take the sequences' phylogenetic relationships into account. We simulated sequences according to realistic evolutionary parameters along real gene trees, and conducted alignments using this re-implementation. We inferred evolutionary rates using both the recently-described method FUBAR (Murrell et al. 2013) and the widely-used PAML M8 model (Yang 2007). Overall, we found that neither the original Guidance filter nor our newly implemented filters confer any significant benefits to positive selection inference. In fact, for alignments with fewer sequences, applying these filters may obscure signal to the point where positive selection inference worsened relative to an unfiltered alignment. In cases where these filters did yield improvements, the magnitude of the benefit was minimal (at most, a 3% increase in true positive rates). Thus, we cannot unequivocally advocate the use of such filters when inferring positive selection in protein-coding sequences.

Results

Guidance re-implementation and analysis pipeline

To systematically evaluate Guidance's influence on positive selection inference, we re-implemented the Guidance software in python and C++. Before conducting any analyses, we verified that, given a set of perturbed alignments, our program produced the same position scores as the original Guidance. In addition to our basic Guidance re-implementation, we derived two new algorithms which employ phylogenetic weighting when assigning confidence scores (see Methods for details). Briefly, the first method incorporated a weight for each sequence in the alignment, as calculated by BranchManager (Stone and Sidow 2007), and the second method incorporated patristic distances (the sum of branch lengths between two taxa). We called these methods, respectively, BMweights and PDweights. We additionally proposed a "gap-penalization score normalization scheme, which naturally assigned lower confidence scores to residues in highly gapped regions, given that such regions were more likely to be poorly aligned. We referred to filters using the gap-penalization scheme as GuidanceP, BMweightsP, and PDweightsP.

Figure **??** shows the overall pipeline for our analysis. First, we simulated

realistic protein-coding sequences using Indelible (Fletcher and Yang 2009) along four different gene trees of sizes 11, 26, 60, and 158 taxa. To ensure that these simulations produced real sequence data to the extent possible, we simulated according to evolutionary parameters inferred from H1N1 influenza hemagluttinin (HA), a protein well known to contain positively selected regions (Meyer and Wilke 2012). We then processed the unaligned sequences with our Guidance re-implementation using the aligner mafft L-INS-I (linsi) (Katoh et al. 2005). We chose to use only linsi, which strongly outperforms other similar alignment softwares for protein alignments (Thompson et al. 2011; Nuin et al. 2006)without sacrificing speed. We scored each alignment using each of our three scoring algorithms (original Guidance, BMweights, and PDweights) with each of our two normalization schemes (original Guidance and gap-penalization), representing a total of six filtering conditions per alignment. We masked residues whose score was below 0.5 with "?".

Finally, we inferred evolutionary rates with both the recently-described FUBAR (Murrell et al. 2013) and PAML's M8 model (Yang 2007) and assessed how alignment filtering affects accuracy in positive selection inference across these two methods. As FUBAR and PAML performed with differing levels of accuracy, we employed different posterior probability cutoffs when assigning sites as positively selected for the two methods. To accomplish this, we used the unfiltered alignments to calculate the posterior probability cutoffs, for each sequence simulation set across inference methods, which yielded an average false positive rate of 1%. This strategy, similar to that undertaken by (Jordan and Goldman 2011), allowed for an unbiased comparison of true positive rates (TPR) between methods. Note that, while we processed all alignments with FUBAR, we did not infer positive selection for the largest simulation set with PAML due to prohibitive runtimes.

To assess the benefits of filtering with a Guidance-based method, we compared the resulting true positive rate (TPR) of positive selection inference between all filtered alignments and their corresponding unfiltered alignment using random-effects linear models for each sequence simulation set, with the simulation count as the random effect to account for the paired structure of our analysis (as in, the same alignment was filtered with multiple strategies, rendering them dependent).

Filtering with Guidance has limited utility

talk about how we did not detect any substantial difference in accuracy among algorithms.

We opted to first infer positive selection with the method FUBAR (Murrell et al. 2013), implemented in HyPhy (Kosakovsky Pond et al. 2005), a very fast alternative to more common evolutionary rate models. textbftransition needed here? I could also mention that the original fubar paper claimed to be more accurate than paml, but then id have to clarify that their assertion was misleading because it's only better than the 3 category paml model. they never compared to the 8 category one.

FUBAR vs PAML

Discussion

all of this is old, probably written early December. Our comprehensive study of alignment filtering with Guidance-based methods indicated that, while masking individual sites rarely hinders positive selection inference, neither does it significantly improve inferences when analyzing protein sequences at realistic divergence levels. These results mirror those made by Jordan and Goldman's (Jordan and Goldman 2011) study on alignment methods and filtering. That the weighted algorithms are unable to improve upon the original Guidance algorithm may indicate the minimal benefits that filtering in this manner produces at all. Were Guidance to offer robust improvements when detecting positively selected sites, one might expect that the more statistically controlled approach would boost the method's performance. However, as we have found that the method itself does not dramatically, if at all, influence positive selection inferences, it is not entirely unexpected that improving the algorithm does not help, either. Based on our results, then, we cannot unequivocally recommend filtering alignments in the manner presented here.

Methods

Guidance Reimplementation

Our reimplemented Guidance is written in Python and C++. Following the algorithm set forth in Penn et al. (Penn et al. 2010), we first create a reference amino-acid alignment using a user-specified progressive alignment software, with choices of clustalw (Thompson et al. 1994), muscle (Edgar 2004), or mafft (Katoh et al. 2002, 2005). For our analysis, we used only mafft L-INS-I (linsi) for all alignments. We then generate 100 bootstrapped alignment replicates, each of which is used to create a bootstrapped tree in FastTree2 (Price et al. 2010). We then use these 100 trees as guide trees in creating 100 new perturbed alignments, which are subsequently compared to the reference alignment to generate a Guidance score for each residue.

Scoring Algorithms

In addition to this basic re-implementation, we implemented two additional scoring algorithms which incorporating phylogenetic information. Before calculating scores, we create a phylogeny using the reference alignment. Our program includes functionality for several maximum likelihood phylogenetic softwares, including FastTree2 (Price et al. 2010) and RAxML (Stamatakis 2006). Using this phylogeny, we can calculate two types of phylogenetic weights. The first uses the software package BranchManager (Stone and Sidow 2007) to calculate a weight for each taxon in the phylogeny representing that taxon's contribution to the phylogeny as a whole. We call this method "BMweights." The second method calculates patristic distances between each taxon in the phylogeny using the python package DendroPy (Sukumaran and Holder 2010). We call this method "PDweights."

We calculated positional confidence scores for each of the n bootstrap alignment as follows. A raw score, S, for a given residue in column j, row i of the reference alignment is calculated as

$$S_{ij} = \sum^{k} s^{j}_{ik} I^{j}_{ik} \tag{1}$$

, where k represents all rows in column j which are not gaps. In this formula,

the indicator function

$$I_{ik}^{j} = \begin{cases} 1 & \text{if reference alignment residue pair (i,k) is present in bootstrap alignment} \\ 0 & \text{if reference alignment residue pair (i,k) is absent in bootstrap alignment} \end{cases} \tag{2}$$

served to compare the bootstrap and reference alignment residue pairings. We calculated s_{ik}^{j} based on the given scoring algorithm:

$$s_{ik}^{j} = \begin{cases} 1 & \text{if Guidance} \\ w_i w_k & \text{if BMweights} \\ d_p(i,k) & \text{if PDweights} \end{cases}, \tag{3}$$

where w_i was the phylogenetic weight of the taxon at row i, as calculated by BranchManager, and $d_p(k,i)$ was the patristic distance (sum of branch lengths) between the taxa at rows i and k.

We then summed positional scores S_{ij} determined from each bootstrap alignment. Scores were normalized to yield a final score $\widetilde{S}_{ij}$ at each residue in the reference alignment using two different schemes. The first scheme, given by

$$\widetilde{S}_{ij} = \frac{S_{ij}}{n \sum_{k}^{k} S_{kj}}, \tag{4}$$

normalized by only considering the residue-residue comparisons made, as done by (Penn et al. 2010). The second scheme was given by

$$\widetilde{S}_{ij} = S_{ij} \Big/ \begin{cases} n \times (l-1) & \text{if Guidance} \\ n \sum d_p(i,l) & \text{if BMweights} \\ n w_i & \text{if PDweights} \end{cases}, \tag{5}$$

where l represented all rows (including gaps) in column j **l is used in two diff ways in the above formula. need better representation!**. This second normalization scheme inherently gave lower scores to highly gapped columns, and we therefore called it the "gap-penalization" normalization.

Thus, in total, we used our Guidance re-implementation to test six different masking algorithms: the original Guidance, weighting using BranchManager weights, and weighting using patristic distance, with two normalization schemes each. We refer to these algorithms as Guidance, BMweights, and PDweights, with their respective gap-penalized versions called GuidanceP, BMweightsP, and PDweightsP.

Sequence Simulation

Coding sequences were simulated using Indelible (Fletcher and Yang 2009). To ensure that our simulations reflect realistic protein sequences, we simulated according to evolutionary parameters of the H1N1 hemagluttinin (HA) influenza protein. To derive these parameters, we aligned 1028 HA protein sequences in mafft linsi (Katoh et al. 2005) of 1038 HA amino acid sequences, and then back-translated to a codon alignment using the original nucleotide sequence data. We generated a phylogeny from this codon alignment in RAxML (Stamatakis 2006) using the "GTRGAMMA" model. Using the codon alignment and phylogeny, we inferred evolutionary parameters with the REL (random effects likelihood) method (Nielsen and Yang 1998) using the software HyPhy (Kosakovsky Pond et al. 2005), with five evolutionary rate categories as free parameters under the GY94 evolutionary model (Goldman and Yang 1994). We employed a Bayes Empirical Bayes approach (Yang et al. 2000) to obtain infer dN/dS values at each site, which we used to assess a complete distribution of site rates. The resulting distribution was log-normal with a mean $dN/dS = 0.37$ with and 8.3% of sites were under positive selection. We binned these rates into 50 equally spaced categories for specification in Indelible, which requires a discrete distribution of dN/dS values. Again according to parameters derived from the HA analysis, kappa was fixed at 5.3 for all simulations. We additionally set the state codon frequencies for our simulations according to those directly calculated from HA alignment.

To simulate across different numbers of taxa, we simulated 100 alignments across four different real gene trees each, yielding a total of 400 simulated alignments. Phylogenies used included an 11-taxon tree of the mammalian olfactory receptor OR5AP2 (Spielman and Wilke 2013), a 26-taxon tree of mammalian rhodopsin sequences(Spielman and Wilke 2013), a 60-sequence tree of phosphoribulokinase (PRK) genes from photosynthetic eukaryotes (Yang et al. 2011), and a 158-taxon multilocus tree of flatfish sequences (Betancur-R et al. 2013). For each simulation set, we directly calculated an indel (insertion-deletion) rates directly from these trees original alignments, to use as simulation parameters, by dividing the total number of gaps present by the total number of positions in each alignment. Respectively, indel rates were 0.053, 0.019, 0.0041, and 0.0066.

Alignment and Positive Selection Inference

Alignments were built with mafft linsi (Katoh et al. 2002, 2005). Using our re-implemented Guidance software, we generated 6 filtered alignments (one for each filtering algorithm), masking residues with scores below cutoff of 0.5. We infered positive selection for every condition using both the recently-described software FUBAR (Murrell et al. 2013) and the widely-used PAML M8 model (Yang 2007). We used the scoring tree, built during the Guidance alignment procedure, as the input phylogeny for selection inference. Therefore, all alignments derived from the same base sequence were processed with identical phylogenies to remove any potential bias. Note that while we employed FUBAR to assess positive selection for all simulation sets, we did not use PAML to infer positive selection for the largest set (158 sequences).

We then compared resulting positive selection inferences for each alignment to its respective true alignment's dN/dS values, given by Indelible during simulation, to assess performance accuracy. As residues may be differently aligned relative to the true simulated alignment, we adopted a consensus method to compare evolutionary rates. In other words, to compare evolutionary rates between between true and inferred alignments, we required that at least 50% of the residues present in a true alignment column be present in an inferred alignment column. If this condition was met, we considered the true alignment columns dN/dS selection status (negative, neutral, or positive) to be the true value for the given inferred column. To assign sites as positively selected, we adopted different posterior probability cutoffs for each simulation set based on the inference method, as FUBAR and PAML have different levels of accuracy. We identified the posterior probability threshold for each simulation set such that the resulting average false positive rate in unfiltered alignments was 1%. The posterior probability cutoffs used are seen in Supplementary **??**.

Statistics were performed using in-house python and R scripts. Linear modeling was conducted using the R packages lme4 (Bates et al. 2012) and multcomp (Hothorn et al. 2008). All code used is available at **the github or something**.

References

Bates D, Maechler M, Bolker B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Betancur-R R, Li C, Munroe T A, Ballesteros J A, Orti G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Systematic Biology 62(5):763–785.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17(4):540–552.

Dwivedi B, Gadagkar S R. 2009. Phylogenetic inference under varying proportions of indel–induced alignment gaps. BMC Evolutionary Biology 9(1):211.

Edgar R C. 2004. Nucleic Acids Research 32:1792–1797.

Fletcher W, Yang Z. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. Molecular Biology and Evolution 26(8):1879–1888.

Fletcher W, Yang Z. 2010. The Effect of Insertions, Deletions, and Alignment Errors on the Branch–Site Test of Positive Selection. Molecular Biology and Evolution 27(10):2257–2267.

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

Hothorn T, Bretz F, Westfall P. 2008. Simultaneous inference in general parametric models. Biometrical Journal 50(3):346–363.

Jordan G, Goldman N. 2011. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29:1125–1139.

Katoh K, Kuma K I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518.

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

Kosakovsky Pond S L, Frost S D W, Muse S V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 12:676–679.

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 Research 21(6):863–874.

Meyer A G, Wilke C O. 2012. Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44.

Murrell B, Moola S, Mabona A, Weighill T, Scheward D, Kosakovsky Pond S L, Scheffler K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Molecular Biology and Evolution 30:1196–1205.

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

Notredame C, Higgins D G, J H. 2000. T–Coffee: a novel method for fast and accurate multiple sequence alignment.

Nuin P A S, Wang Z, Tillier E R M. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471.

Ogden T H, Rosenberg M S. 2006. Multiple Sequence Alignment Accuracy and Phylogenetic Inference. Systematic Biology 55(2):314–328.

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

Price M N, Dehal P S, Arkin A P. 2010. FastTree2: Approximately maximum–likelihood trees for large alignments. PLoS One 5:e9490.

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.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet G H, Graur D. 2009. Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biology and Evolution 1(0):114–118.

Spielman S J, Wilke C O. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein–coupled receptors. Journal of Molecular Evolution 76:172–182.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:21:2688–2690.

Stone E, Sidow A. 2007. Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics 8:222.

Sukumaran J, Holder M T. 2010. DendroPy: A python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56(4):564–577.

Thompson J D, Higgins D G, Gibson T J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position–specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673–4680.

Thompson J D, Linard B, Lecompte O, Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. PLoS ONE 6(3):e18093.

Yang Y, Maruyama S, Sekimoto H, Sakayama H, Nozaki H. 2011. An extended phylogenetic analysis reveals ancient origin of "non-green" phosphoribulokinase genes from two lineages of "green" secondary photosynthetic eukaryotes: Euglenophyta and Chlorarachniophyta. BMC Research Notes 4:330.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.

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

Table 1: Effect of alignment filtering on the true positive rate of positive selection inference.

Simulation Set	Method	Unfiltered	GuidanceP	BMweightsP	PDweightsP
11 taxa	Fubar	0.108	0.109	0.110	0.110
	Paml	0.0873	0.0878	0.0881	0.0880
26 taxa	Fubar	0.229	0.232	0.233	0.233
	Paml	0.194	**0.204***	**0.203***	**0.203***
60 taxa	Fubar	0.345	**0.379****	**0.377****	**0.375****
	Paml	0.447	0.447	0.440	0.445
158 taxa	Fubar	0.374	**0.388****	**0.387****	**0.387****

NOTE.— Significance levels: $^{**}P < 1 \times 10^{-5}$; $^{*}P < 1 \times 10^{-4}$. *Unfiltered*: average true positive rate (TPR) for unfiltered alignments; *GuidanceP*: average true positive rate (TPR) for alignments filtered with GuidanceP algorithm; *BMweightsP*: average TPR for alignments filtered with BMweightsP algorithm; *PDweightsP*: average TPR for alignments filtered with PDweightsP algorithm. All significance levels are relative to the given simulation set's unfiltered alignment. We detected no significant TPR differences among filters tested within sequence simulation sets. All significance levels were corrected for multiple comparisons with the single-step method.