(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

People are really into masking alignments these days. Active development of methods. Here, we comprehensively test one common masking method and show that it's probably not as awesome as you thought. Is it bad? Probably not. But is it going to make or break you? Also no. Is it worth your time? Doubtful. Plus the actual guidance is pretty damn buggy, and that's ridiculous.

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, the elevated false positive rates incurred during positive selection inference when using an alignment of poor quality has been a troubling observation (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 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.

One widely used **(should i cite some papers that use guidance?)** software for alignment filtering is Guidance (Penn et al. 2010). Guidance bootstraps alignments by sampling variants in the alignment guide trees. Ultimately, Guidance derives a confidence score for each position in the alignment (see Penn et al. for details). Users can then "mask" (i.e. replace with an ambiguous character such as "?") positions with scores below a certain 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) or indel rates (0.2 indel/substitution events), Jordan and Goldman found that Guidance significantly boosted true positive rates. While these results are compelling, protein sequences used in positive selection studies, however, rarely, if ever, contain sequences separated by such high divergences; the MPL for an average mammalian gene

tree is only about 0.05 **citation needed maybe?**, or $1/36th$ 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 worsens relative to an unfiltered alignment. In cases where these filters do yield improvements, the magnitude of the benefit is 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

To systematically evaluate Guidance's influence on positive selection inference, we re-implemented the Guidance software. This re-implementation includes two new scoring algorithms which employ phylogenetic weighting when assigning residue scores (see Methods for details). Briefly, the first method of phylogenetic weighting incorporates individual taxon weights calculated by BranchManager (Stone and Sidow 2007), and the second method incorporates patristic distances, that is, the sum of branch lengths between two taxa. We call these methods, respectively, BMweights and PDweights. We additionally propose a gap-penalization score normalization scheme, which naturally assigns lower confidence scores to residues in highly gapped regions. Given that residues in such regions are more likely to be poorly aligned than residues elsewhere in the MSA, this strategy imposes a more strict scoring threshold for such residues. Each of the three scoring algorithms (original Guidance, BMweights, and PDweights) was conducted with both regular normalization, as conducted in the original Guidance implementation (Penn et al. 2010) and gap-penalized normalization, representing a total of six masking algorithms per alignment.

Figure 1 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 opted to use only linsi for all alignments, as recent studies have demonstrated that this aligner strongly outperforms other similar alignment softwares for protein alignments

(Thompson et al. 2011; Nuin et al. 2006) without sacrificing speed. After calculating confidence scores at each alignment position, we had to select a score threshold below which to mask scores. Preliminary analyses (see Methods for details) indicated that the optimal cutoff for gap-penalized algorithms was 0.5 and for regular algorithms was 0.7, so we masked alignments accordingly. Finally, we inferred evolutionary rates with both FUBAR (Murrell et al. 2013) and PAML's M8 model (Yang 2007) and assessed how alignment filtering affects accuracy in positive selection inference. We interpreted sites to be under positive selection if the given inference method returned a posterior probability greater than or equal to 0.9. Note that, while we processed all alignments with FUBAR, but we did not use PAML to infer positive selection for the largest alignments (158 taxa) due to prohibitive runtimes.

To assess the benefits of filtering with a Guidance-based method, we compared the true positive rates between all filtered alignments (processed with each of six scoring algorithms) and their corresponding unfiltered, or reference, alignment. For each sequence simulation set, we built a random-effects linear model to compare performances across scoring schemes. These models consisted of either the true positive or false positive rate as the response variable, with the filtering algorithm as a fixed effect and simulation as a random effect.

Interestingly, we found that masking does not produce the same effect when positive selection is inferred with FUBAR or PAML. Specifically, filtering residues tended to increase accuracy more with FUBAR than with PAML. This apparent discrepancy likely reflects algorithmic differences between the two methods; while FUBAR simply approximates whether each site in the alignment is under positive selection or not, PAML yields precise dN/dS point estimates at each site. It seems, then, that missing information influences approximate methods more significantly than precise ones.

Further, of all filters tested, no one masking algorithm clearly stood out as more or less accurate than any other. All filters applied yielded statistically similar performances when used in positive selection inference. Thus, incorporating phylogenetic weighting into the Guidance scoring algorithm, and filtering residues accordingly, did not seem to yield a statistical benefit in positive selection inference.

Overall, we found that filtering with Guidance or a Guidance-based method yields modest, if any, benefits under realistic simulation conditions. For the 11-sequence simulation set, neither FUBAR nor PAML showed any significant difference in true positive rates between the unfiltered and any reference alignment. Similarly, we found no difference between FUBARs positive selection inferences between any filtered alignment and the unfiltered alignment for the 26-sequence data set. Analysis with PAML did boost, albeit at most by 0.97%, true positive rates for filtered alignments with this 26-sequence simulation set. Even so, not filtering and using FUBAR yielded more accurate results than did filtering and then inferring positive selection with PAML. FUBAR did significantly increase true positive rates for the 60-sequence simulation set, although the magnitude of this effect was, at most, 3%. Figure 2 displays an ROC curve for FUBAR's performance in analyzing this simulation set, demonstrating that, while significant, effects of masking are rather minimal. Oddly, while this data set yielded the most robust improvements in filtered alignments relative to the reference alignments when processed with FUBAR, processing with PAML does not reveal any significant difference between any masking algorithm and the reference alignment. The 158-sequence simulation set, when analyzed with FUBAR, similarly showed that masking

can yield significant but minimal improvements, with increases in true positive rate on the order of roughly 1-2%.

Similar to Jordan and Goldman's study (Jordan and Goldman 2011), we detected extremely low false positive rates across all alignments, ranging from an average of 0.09% to 0.2%. While we did detect some significant differences in false positive rate among certain filtering conditions, these differences were of nearly negligible magnitude. When processed with FUBAR, filtered alignments for the 11-sequence simulation set showed roughly a 0.01% reduction in false positive rate, but we found no significant difference in false positive rates between filtered and unfiltered alignments for the either the 26-sequence or the 158-sequence simulation sets. Interestingly, FUBAR's results for the 60-sequence simulation set demonstrated increased false positive rates in the filtered alignments relative to the unfiltered alignment, at an average magnitude of 0.03%. Analysis with PAML did revealed no significant differences in false positive rates between filtered or unfiltered alignments for the 11-sequence and the 26-sequence simulation sets. Alternatively, PAML's results for the 60-sequence simulation set did show significant decreases false positive rates for all filters relative to the reference alignment, ranging from 0.01% to 0.07% decreases. In sum, although there are some significant differences in false positive rates between filtered and unfiltered alignments, the magnitudes are exceedingly minimal. Filtering, then, may decrease false positive inferences under certain conditions, but to a negligible extent.

Discussion

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 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." The original Guidance (see (Penn et al. 2010) for a more detailed explanation), assigns each residue-residue pair in column of a perturbed alignment column a score of 1 or 0, based on whether those residues are also paired, in the reference alignment. We substitute this calculation with our two phylogenetic weighting schemes as follows.For the BranchManager-based weighting scheme, we score correctly paired residues as the product of their phylogenetic weights, and incorrectly paired residues receive a score of 0. For the patristic distance weighting scheme, we score correctly paired residues as the patristic distance between them, and again incorrectly paired residues receive a score of 0. The original Guidance performs this scoring for each residue and normalizes over the total number of comparisons made, rather than the total number of sequences in the alignment. Using this method, gapped sites are not considered during scoring. However, it is also possible to consider gapped sites in normalizing residue scores. Thus, we normalize scores in two ways: first by normalizing based on the the number of residue-residue comparisons made, and second normalizing based on the total number of sequences. We refer to the second normalization method as a "gap-penalization" scheme, as sites in highly gapped columns will inherently receive lower scores. Thus, in total, we use 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 generated using our re-implemented Guidance and weighted guidance using "mafft –auto," which automatically selects either the linsi/einsi/ginsi algorithm for a given alignment based on its properties. From each simulated dataset, we saved filtered alignments, with residues masked at the four different cutoffs 0.3, 0.5, 0.7, and 0.9 across all scoring algorithms, as well as an unfiltered reference alignment. We assessed 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) due to prohibitive runtime.

We then compared resulting positive selection inferences for each alignment to their respective true alignment's dN/dS values, given by Indelible during simulation to assess performance among filters. As residues may be differently aligned relative to the true sim-

ulated 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 omega value to be the true value for the given inferred column.

We determined the optimal masking threshold (of 0.3, 0.5, 0.7, and 0.9) for each scoring algorithm according to a procedure laid out by Jordan and Goldman (Jordan and Goldman 2011); we assessed true positive rates for each alignment where the false positive rate was roughly 1%. This method controls for accuracy differences between positive selection inference methods. Comparing true positive rates at a fixed false positive rate of 1% demonstrated that, for gap-penalization algorithms, cutoffs of 0.3 and 0.5 perform the best, and for the regular algorithms, a cutoff of 0.7 performs the best. We chose to use cutoffs of 0.5 and 0.7 for the gap-penalized and regular scoring procedures, respectively.

All was analyzed with in-house python and R scripts, *available at the github or something.*

References

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.

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

Nuin P A, Wang Z, Tillier E R. 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, Wilke C. 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, Higgins D, Gibson T. 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.

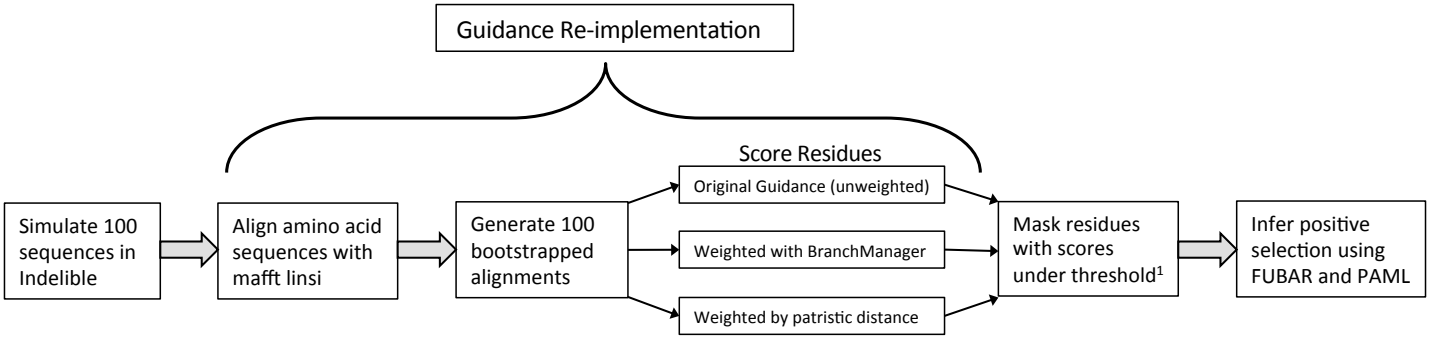

Figure 1: INSERT CAPTION HERE.

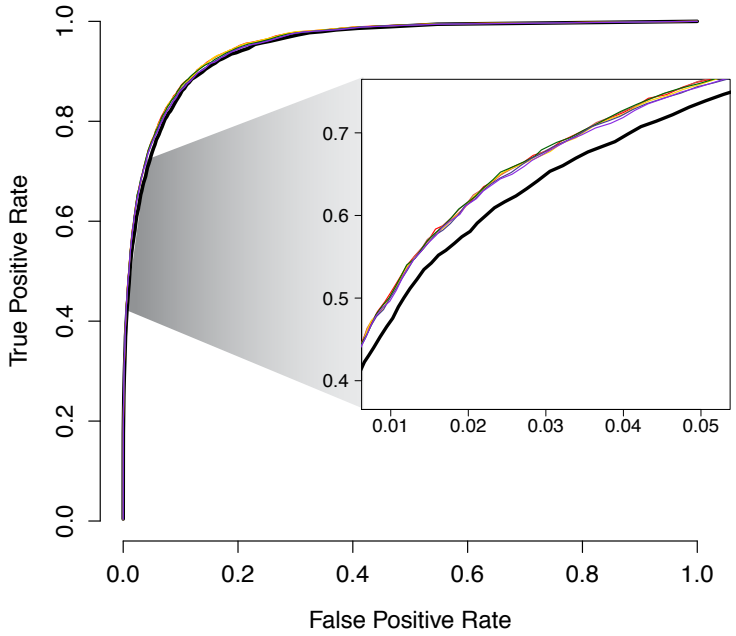

Figure 2: INSERT CAPTION HERE. No algorithm clearly stands out. Roughly 3% above the reference alignment curve. So minimal!