Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions

Eyal Privman,†[,1,2,3] Osnat Penn,†[,1] and Tal Pupko*[,1,4]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

[2]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]National Evolutionary Synthesis Center, Durham, North Carolina

†These authors contributed equally to this work

***Corresponding author**: E-mail: talp@post.tau.ac.il.

Associate editor: Koichiro Tamura

Abstract

Errors in the inferred multiple sequence alignment may lead to false prediction of positive selection. Recently, methods for detecting unreliable alignment regions were developed and were shown to accurately identify incorrectly aligned regions. While removing unreliable alignment regions is expected to increase the accuracy of positive selection inference, such filtering may also significantly decrease the power of the test, as positively selected regions are fast evolving, and those same regions are often those that are difficult to align. Here, we used realistic simulations that mimic sequence evolution of HIV-1 genes to test the hypothesis that the performance of positive selection inference using codon models can be improved by removing unreliable alignment regions. Our study shows that the benefit of removing unreliable regions exceeds the loss of power due to the removal of some of the true positively selected sites.

Key words: multiple sequence alignment, GUIDANCE, alignment reliability, positive selection, molecular evolution, phylogeny.

Multiple sequence alignments (MSAs) contain many unreliable regions that may lead to false inference of positive selection (Wong et al. 2008; Schneider et al. 2009; Fletcher and Yang 2010). It thus seems obvious, at first glance, that such regions should simply be removed from the MSA before any test for positive selection is applied. However, positively selected sites are fast evolving, often displaying high rates of both substitutions and insertions/deletions (indels). Therefore, fast evolving sites are likely to be found in unreliably aligned regions, and it may be that by removing such regions, the power of tests for detecting positive selection would be much diminished. In this study, we used simulations to study the impact of removing unreliable alignment regions on positive selection inference. We demonstrate that the benefit from removing unreliable regions exceeds that of retaining them, both for the task of testing whether a gene has experienced positive selection and for the task of identifying specific positively selected sites.

We simulated 150 data sets of 20 protein-coding sequences in a way that mimics the evolution of 3 genes of the HIV-1: *gag*, *pol*, and *env* (see Materials and Methods). HIV-1 was chosen as a classical example for fast evolving sequences with extensive positive selection due to the interaction with the host immune system (Rambaut et al. 2004). The *env* gene codes for the envelope protein, which contains the major targets for antibodies of the immune system. The hypervariable regions in *env* sequences (Holmes et al. 1999, p. 196) are often challenging for alignment algorithms. Conversely, *pol* codes for intracellular enzymes whose structure and catalytic sites must be conserved to maintain their function. The *gag* gene codes for intracellular structural components and therefore displays an intermediate level of sequence variation (see results below).

The simulated data sets were aligned using either CLUSTALW (Thompson et al. 1994), MUSCLE (Edgar 2004), MAFFT (Katoh et al. 2005), or PRANK (Loytynoja and Goldman 2005). We ran PAML (Yang 2007) to calculate the likelihood ratio test (LRT) score for the entire gene having experienced positive selection and recorded site-specific posterior probability scores. The knowledge of the sites that were simulated with positive selection allows us to evaluate the accuracy and power of inference, using receiver operating characteristic (ROC) analysis with the PAML site-specific recorded score as the predictor (Green and Swets 1966; Fawcett 2006). In these ROC curves, a false-positive (FP) prediction, for example, is an alignment column that was simulated without positive selection, but PAML inferred it to be positively selected. ROC analysis was performed before and after the removal of unreliable regions. Unreliable regions should ideally reflect alignment errors. Hence, we began by testing the impact of removing true alignment errors, known based on the simulation.

Example ROC curves before (in blue) and after (in red) filtering true alignment errors are plotted in figure 1*a*, based on MAFFT alignments of simulated *gag*-like sequences. In this scenario, filtering improves the true positive (TP) rate for an FP rate of less than 15% but not for higher FP rates. We concentrate on the lower FP rates because most

Letter

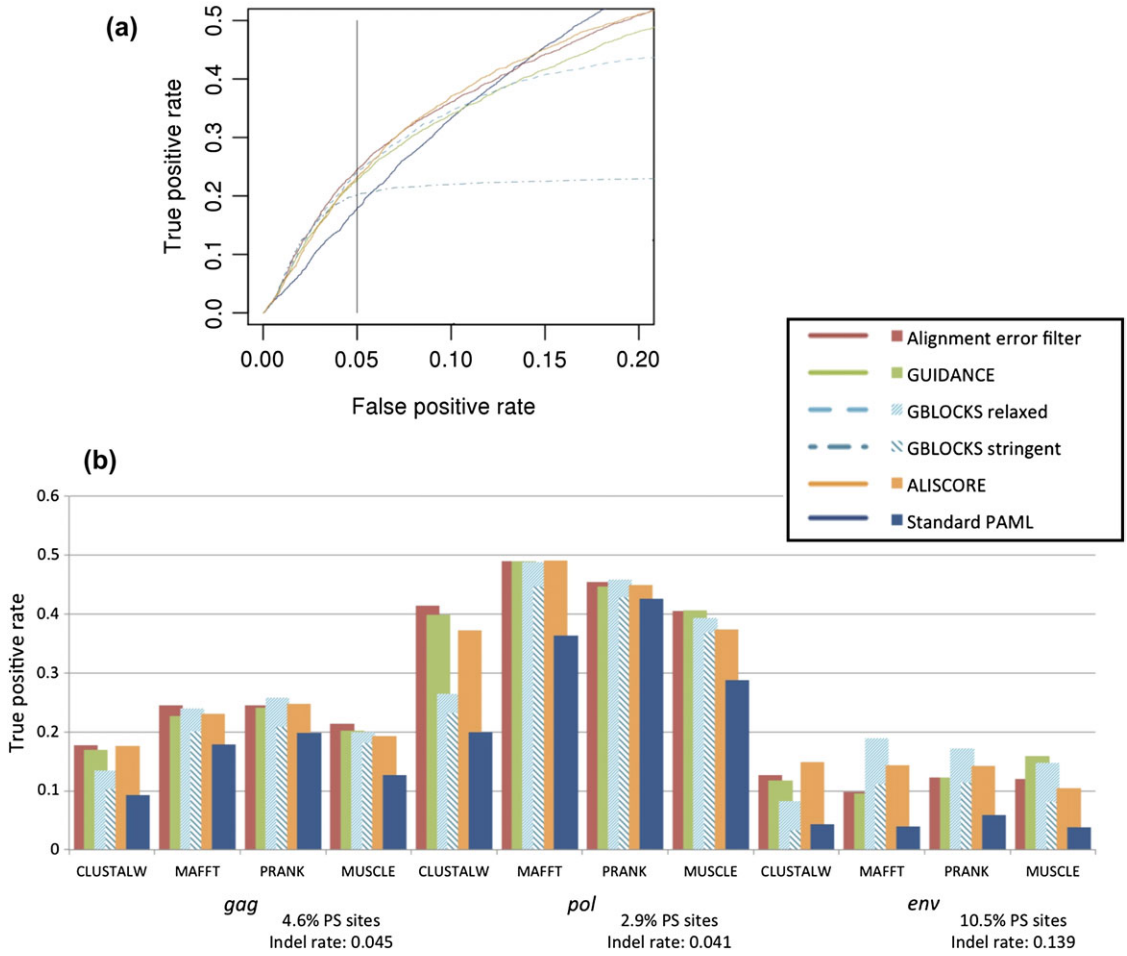

Fig. 1. Accuracy and power of positive selection inference. (*a*) ROC plots for the sensitivity and specificity of PAML predictions compared with predictions after filtering columns with true alignment errors or using GUIDANCE, GBLOCKS, or ALISCORE filtering. The vertical line marks an FP rate of 0.05. Showing results for 50 replicates of MAFFT alignments for simulations based on the *gag* gene. Only values between 0 and 0.2 FP rates are presented to emphasize the desired range. (*b*) The TP rate when the FP rate is 0.05 for standard PAML (full blue) compared with filtered results as in (*a*). Showing results for 50 replicates of simulations based on each of the *gag*, *pol*, or *env* genes (the proportion of sites simulated with positive selection (PS sites) and the average indel rate is specified for each). Sequences were aligned by CLUSTALW, MUSCLE, MAFFT, or PRANK.

researchers are typically interested in an FP rate of 5%. In figure 1*b*, we fixed the FP rate at 0.05 and compared the TP rate for the standard PAML run with the rate obtained after filtering, for the *gag*, *pol*, or *env* simulations and using MAFFT, MUSCLE, CLUSTALW, or PRANK. The full set of ROC curves is included as supplementary figures S3–S14, Supplementary Material online. These results demonstrate that filtering erroneous alignment columns improves the performance of site-specific positive selection inference. Note that *env*-like simulations are the most challenging, as expected for the higher proportion of positively selected sites and the higher indel rates (see bottom of fig. 1*b*; supplementary fig. S2, Supplementary Material online). Note also that the performance of positive selection identification depends on the MSA algorithm. Aligning with PRANK leads to the best performance, whereas CLUSTALW to the worst. The greatest contribution of filtering is evident in the lower scoring alignments. In contrast, the contribution of filtering is smaller where the performance was already high before filtering (e.g., PRANK alignments for *pol* simulations).

Second, we filtered unreliable columns inferred by either of the alignment confidence methods GUIDANCE (Penn et al. 2010b), ALISCORE (Misof and Misof 2009), and GBLOCKS (Talavera and Castresana 2007) (see Materials and Methods). All filtering methods achieve some improvement over standard PAML runs (fig. 1*a* and *b*). We observe only small differences in performance between the filtering of the true errors and the filtering based on alignment confidence estimates. This result demonstrates that confidence scores can achieve almost the full contribution possible from the knowledge of the actual alignment errors and suggests that the signal lost due to the removal of sites that genuinely evolved under positive selection is comparable to the case when the true alignment is known.

Third, to minimize information lost in the filtering of whole columns, we attempted to filter individual unreliably aligned residues using the GUIDANCE residue-specific score (Penn et al. 2010a). Because it is not possible to remove a residue from an MSA, we instead masked unreliably aligned codons (with scores < 0.9) by replacing them with

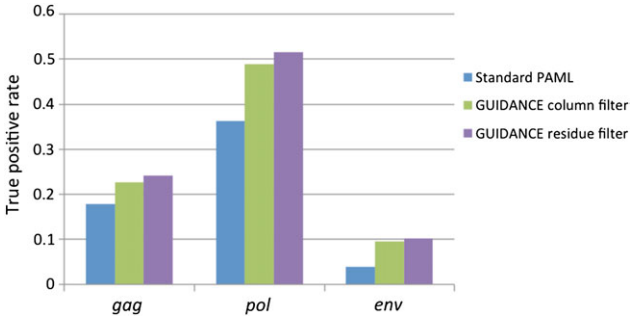

FIG. 2. Column versus residue filtering. A plot of the TP rate when the FP rate is 0.05. Filtering of unreliable columns (as in fig. 1) compared with the filtering of unreliable residues (purple). Each data point represents 50 simulation replicates aligned using MAFFT.

a missing data character, i.e., an "NNN" codon. This approach further improved the performance compared with the above-mentioned filtering of columns (fig. 2). Thus, masking specific residues proves to be a more effective strategy to eliminate false inference due to alignment errors without unnecessary loss of well-aligned residues.

Forth, we tested whether filtering also improves the performance of the identification of whole genes as evolving under positive selection. That is, improving the performance of the LRT (see Materials and Methods). Toward this goal, in addition to the 150 data sets used in the above analyses, we simulated 150 data sets under similar conditions but without positive selection. Figure 3 shows the average FP and FN rates for simulated *gag*, *pol*, and *env* data sets aligned by MAFFT. As expected, GUIDANCE filtering reduces the FP error rate and increases the FN rate. The FN rate for the residue filter is lower than the column filter. In all three scenarios, the FP rate reduced to almost 0. Notably, the most dramatic effect is evident from the *env* data. Here, we do not find the problem of high FN rate, which is not surprising for data simulated with more positive selection. However, the FP rate was 90%, which we attribute to the larger indel rates. GUIDANCE filtering reduces the FP rate to 0, while the FN rate remains 0. To conclude this point, filtering unreliable alignment regions as determined by GUIDANCE is also valuable for the task of determining whether or not a protein evolves under positive selection, although some scenario may suffer from a substantial increase in the FN rate.

Our results can be viewed through three cross-sections: First, a clear trend relates to the choice of alignment algorithm. In terms of the accuracy of prediction before filtering, PRANK consistently outperforms the other aligners, CLUSTALW generally scores lowest, and MAFFT and MUSCLE are intermediate. This may be attributed to the correction in PRANK of the bias toward overalignment, which is inherent in the common implementation of progressive alignment. Loytynoja and Goldman (2008) demonstrated that state-of-the-art MSA programs (including CLUSTALW, MUSCLE, and MAFFT) are biased toward producing too short an alignment compared with the true solution, whereas PRANK is almost unbiased. Overalignment essentially means aligning in the same column residues or codons that are not truly homologous. Columns containing overalignment errors may be responsible to the increased FP rate of positive selection inference. Also consistent with this explanation for our results is their finding that CLUSTALW is the most biased algorithm toward overalignment, whereas MAFFT and MUSCLE are approximately midway between CLUSTALW and PRANK. Other recent studies also report an advantage for PRANK in the context of positive selection inference (Fletcher and Yang 2010; Markova-Raina and Petrov 2011).

Second, our results compare three different measures of alignment confidence. In this dimension, we do not observe a clear advantage for any of the methods. A different result may be found in other simulation scenarios, as previous studies argued that GUIDANCE and ALISCORE are superior to GBLOCKS in the detection of alignment errors (Kuck et al. 2010; Penn et al. 2010b). In the comparison between GUIDANCE column and residue filters, we did find a small advantage to the residue filter, which we attribute to the reduced loss of information. Notably, residue scores are only implemented in GUIDANCE.

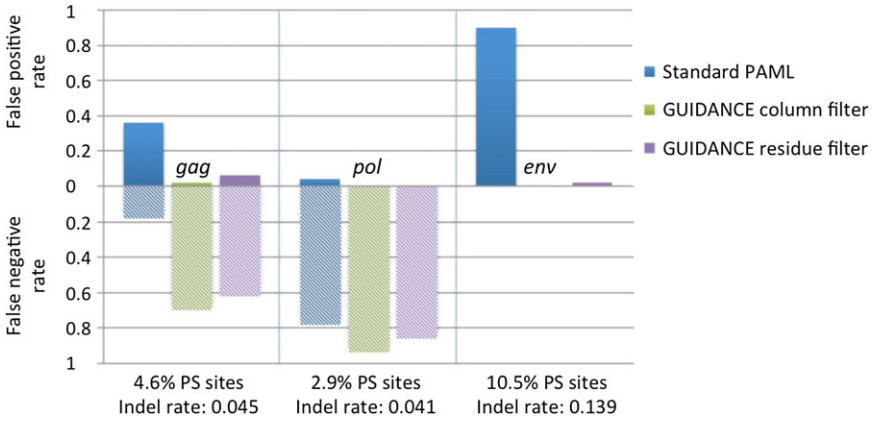

FIG. 3. False-positive and false-negative rates in the LRT for whole genes. Error rates were calculated for 50 replicates of simulations with and without positive selection, for each of the *gag*, *pol*, and *env* gene models, aligned by MAFFT. Results of standard PAML runs are compared with PAML runs on alignments filtered by the GUIDANCE column score or residue score. The proportion of sites simulated with positive selection (PS sites) and the average indel rate is written for each gene model.

A third trend is that the poorest performance was found in *env*-like simulations and the best was found in *pol*-like simulations. Accordingly, the greatest improvement after filtering was found in the *env* simulations. This may be attributed to the above-mentioned distinct functional roles of these genes that make them subject to dramatically different selection constraints and pressures. The increase in LRT false negatives due to filtering also varies substantially among the three genes (fig. 3). This raises the question if filtering should be applied for any data set analyzed prior to positive selection inference. Without filtering, our results show that the FP rate can be as high as 90%, depending on the data set in question. Researchers cannot know what FP rate to expect for their data. It is not under control. Conversely, we show that in all three scenarios, positive selection can still be detected after filtering. Therefore, filtering provides a more conservative approach to positive selection inference, and we believe that the filtering of unreliable regions should indeed be applied prior to any positive selection analysis. Clearly, the development of better methods for alignment filtering and methods that can account for alignment uncertainty within the inference scheme of positive selection may contribute further to this issue.

Materials and Methods

We simulated 150 data sets using INDELible (Fletcher and Yang 2009), which is capable of simulating codon sequences using models of both site-specific positive selection and indel events. We took special care to design a realistic simulation scheme that captures the characteristics of naturally evolving sequences, accounting for spatial variation of both indel events and K_a/K_s values across the alignment. Thus, we first inferred relevant indel and K_a/K_s parameters from coding *gag*, *pol*, and *env* sequences and then simulated sequences according to these parameters. Detailed description of the simulation scheme is given as supplementary text, Supplementary Material online.

Alignment was performed on the translated sequences and then reverse-translated to get codon alignments (implemented in the GUIDANCE program). PAML (version 4.4) was run for the M8 and M8a models, using the Bayes empirical Bayes method, and with the flag cleandata = 0. LRT was calculated to determine if M8 fits the data better than M8a, using a *p*-value threshold of 0.05 (supplementary table S1, Supplementary Material online, lists the number of data sets that passed the test). Comparison of site-specific positive selection inference was computed only on those alignments for which the LRT indicates the presence of positive selection. This mimics the standard in the field, in which site-specific inference is only done in cases where the entire gene shows support for positive selection. Filtering of columns was done by resetting their PAML posterior probability score to zero, effectively turning a positive prediction into a negative one. To mask individual unreliably aligned residues, we replaced their codons with "NNN". GUIDANCE was run with 100 bootstrap iterations. We filtered columns with scores less than 99% and residues with scores less than 90%. GBLOCKS was run with either "relaxed" or "stringent" settings as defined by Talavera and Castresana (2007). ALISCORE was run with default parameters. ROC analyses were conducted using the ROCR package (Sing et al. 2005).

Supplementary Material

Supplementary text, table S1, and figures S1–S14 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/)

Acknowledgements

We thank Itay Mayrose and Nimrod Rubinstein for comments and suggestions. This study was supported by an Israel Science Foundation grant 878/09 and by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606 to T.P. O.P. was supported by the Edmond J. Safra Bioinformatics program at Tel-Aviv University and by the Converging Technologies scholarship program.

References

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

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognit Lett.* 27:861–874.

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.

Green DM, Swets JA. 1966. Signal detection theory and phycophysics. New York: John Wiley & Sons.

Holmes E, Pybus O, Harvey P. 1999. The molecular population dynamics of HIV-I. In: Crandall K, editor. The evolution of HIV Balimore, (MD): Johns Hopkins University Press. p. 177–203.

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

Kuck P, Meusemann K, Dambach J, Thormann B, von Reumont B, Wagele J, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology* 7:10.

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

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

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.

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.

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 (suppl):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.

Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. *Nat Rev Genet.* 5:52–61.

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

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*. 21:3940–3941.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

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

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