

Sequence analysis

Combining statistical alignment and phylogenetic footprinting to detect regulatory elements

Rahul Satija^{1,*}, Lior Pachter² and Jotun Hein¹¹Department of Statistics, Oxford University, Oxford, UK and ²Department of Mathematics, University of California, Berkeley, CA, USA

Received on January 21, 2008; revised on February 21, 2008; accepted on March 17, 2008

Advance Access publication March 18, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Traditional alignment-based phylogenetic footprinting approaches make predictions on the basis of a single assumed alignment. The predictions are therefore highly sensitive to alignment errors or regions of alignment uncertainty. Alternatively, statistical alignment methods provide a framework for performing phylogenetic analyses by examining a distribution of alignments.

Results: We developed a novel algorithm for predicting functional elements by combining statistical alignment and phylogenetic footprinting (SAPF). SAPF simultaneously performs both alignment and annotation by combining phylogenetic footprinting techniques with a hidden Markov model (HMM) transducer-based multiple alignment model, and can analyze sequence data from multiple sequences. We assessed SAPF's predictive performance on two simulated datasets and three well-annotated *cis*-regulatory modules from newly sequenced *Drosophila* genomes. The results demonstrate that removing the traditional dependence on a single alignment can significantly augment the predictive performance, especially when there is uncertainty in the alignment of functional regions.

Availability: SAPF is freely available to download online at <http://www.stats.ox.ac.uk/~satija/SAPF/>

Contact: satija@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The annotation of *cis*-regulatory regions to identify functional regulatory elements remains a difficult and important problem in computational biology. Phylogenetic footprinting (Tagle *et al.*, 1988) approaches assume that regulatory elements in non-coding regions are subject to purifying selection, and therefore will exhibit higher levels of conservation than surrounding neutral sequence. Numerous phylogenetic footprinting approaches have been developed and successfully applied to detect conserved regulatory elements in diverse taxa (Boffelli *et al.*, 2003; Cliften *et al.*, 2003; Stark *et al.*, 2007). A popular approach used in the creation of University of California Santa-Cruz (UCSC) Genome Browser conservation scores,

phastCons, implements a hidden Markov model (HMM) with a hidden state for conserved regions, and a hidden state for non-conserved regions (Siepel *et al.*, 2005). Conserved elements are predicted by fitting the HMM to an alignment by maximum likelihood. The algorithm, however, assumes a single 'perfect' alignment, and thus the predicted annotations are highly dependent on the accuracy of the alignment. This dependence can have significantly deleterious effects on the accuracy of the predictions. Stark *et al.* (2007) analyzed 12 *Drosophila* genomes for *de novo* discovery of functional elements. When investigating the effect of alignment accuracy on their predictions, they found only 59% similarity between different alignment strategies for regulatory motif instances. Additionally, a simulation study by Pollard *et al.* (2006) found that variation in alignment accuracy could cause significant errors in evolutionary studies, and discussed the need for phylogenetic tools that can control for alignment errors.

Comparative approaches that do not assume a single alignment have been previously implemented to predict regulatory elements from two-species comparisons. Wasserman *et al.* (2000) analyzed orthologous sequences from the human and mouse genomes by analyzing a distribution of alignments using the Bayes Block Aligner (Zhu, 1998), although the algorithm did not consider separate models for regulatory elements and background sequence. Sinha and He (2007) developed MORPH, a framework which detects and aligns instances of known motifs by summing over all possible alignments of two species. While this approach has the additional requirement that binding site motifs must be known in advance, the authors report that binding-site predictions are robust to alignment ambiguities.

Statistical alignment is a probabilistic approach in which explicit models for evolutionary events (substitutions, insertions, deletions) are applied to the data in order to calculate not only the maximum-likelihood alignment, but in addition, the probability distribution for all alignments. One of the first models, TKF91, proposed a time-reversible evolutionary model treating insertions and deletions (indels) as single-nucleotide events (Thorne *et al.*, 1991). Future work enabled the modeling of indel events of multiple nucleotide fragments with geometrically distributed lengths, and models which can assume an arbitrary distribution on the lengths of indel events (Miklós *et al.*, 2004; Thorne *et al.*, 1992; Wang *et al.*, 2006). These models provide a framework for statistical evolutionary inference without

*To whom correspondence should be addressed.

assuming a single alignment. For example, model parameters specifying the frequency and length of evolutionary events can be estimated by maximum likelihood from a probability-weighted distribution of alignments (Holmes, 2005; Holmes and Rubin, 2002), and a recent study demonstrated that single alignments dramatically underestimate the rate of indels between human and mouse (Lunter, 2007).

We created SAPF, a statistical aligner and phylogenetic footprinter, a new program for annotating regulatory elements. SAPF is currently the only computational tool with the ability to do all of the following:

- (1) Distinguish between sequence undergoing neutral evolution and conserved sequence that may be the result of purifying selection.
- (2) Make these predictions by calculating and summing over a distribution of many possible alignments, instead of a single assumed alignment.
- (3) Analyze sequence data from multiple organisms, related by any previously known phylogenetic tree.
- (4) Assume that insertion and deletion events have a geometrically distributed length, with a distribution parameter that can be set in the model.

2 METHODS

In order to infer the locations of functional elements without assuming a single alignment, SAPF combines the features of a statistical aligner and a phylogenetic footprinter. Both alignment and annotation are performed by a HMM, referred to as the SAPF HMM. A path of SAPF HMM hidden states that traverses a set of input sequences represents a multiple sequence alignment as well as an annotation of each alignment column as either functional or neutral sequence. Thus, by calculating the likelihoods of many different state paths, SAPF can infer the locations of functional elements based on a probability-weighted distribution of alignments.

SAPF implements a statistical alignment model composed from individual HMM transducers. An HMM transducer models the evolution of an ancestral sequence into a descendent sequence, and thus models an evolutionary process on a single branch of a phylogenetic tree. Combining different transducers together creates a multiple sequence hidden Markov model (MHMM) which allows SAPF to model an evolutionary process on an entire tree and to infer a multiple sequence alignment. In the following subsections, we describe the characteristics and composition of this MHMM, demonstrate how we adapt the model to include phylogenetic footprinting capabilities, and show how the SAPF HMM can be used to perform alignment-free inference of functional elements.

2.1 The statistical alignment MHMM

The multiple sequence HMM (MHMM) used for statistical alignment is composed through the combination of individual HMM transducers. Finite-state transducers have recently emerged as a promising framework for appropriately modeling the evolution of insertion and deletion mutations (Bradley and Holmes, 2007; Holmes, 2003). A transducer is very similar to a traditional pairwise HMM (Durbin *et al.*, 1998). A key difference, however, is that a pairwise HMM models two sequences evolving from a common ancestor, while a transducer models the ancestral sequence evolving into the descendent sequence.

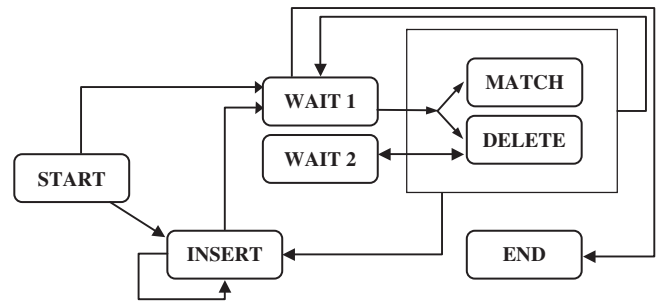


Fig. 1. Branch HMM representing evolutionary model with geometrically distributed indel lengths. The second wait state allows the delete state to effectively self-transition.

The ability to compose multiple transducers in either serial or parallel makes the transducer framework ideal for phylogenetic sequence analysis, since each branch of the phylogenetic tree can be represented by a transducer, referred to as a branch HMM. In order for individual transducers to be combined, Holmes (2003) introduces a type of null (non-emitting) state known as a wait state. When the transducer is in a wait state, it is waiting for an input letter from the ancestral sequence, which must be received before the next state transition can occur.

We modeled an evolutionary process that allows for insertion and deletion events of a geometrically distributed length, so we created a branch HMM to fit these requirements. The model is represented graphically in Figure 1, with arrows representing potential state transitions. Insertion lengths are geometrically distributed since the insertion state has the ability to self-transition, and the parameter for this geometric distribution is set by the self-transition probability for the insert state. We added a second wait state since the deletion state is unable to self-transition. By adding a second wait state that only a delete state can transition to, making it effectively a ‘within a deletion’ state, the transducer models geometrically distributed deletions as well. See the HMM Transducers section in the Supplementary Material for a more complete discussion of transducer theory, the wait state, and the branch HMM.

Ian Holmes has published an exact algorithm demonstrating how to combine a phylogenetic tree with branch HMMs in order to create a combined model capable of analyzing multiple sequence data (Holmes, 2003). Holmes refers to this overall model as a multiple-sequence HMM (MHMM), and generously provides the open-source program phyloComposer which generates the state space for these multiple-sequence models, as well as their associated transition probabilities (Holmes, 2007). The MHMM generates symbols from the ancestral root node of the tree and then propagates these symbols down through each of the branches. For a full discussion of the MHMM state space, transition and emission probabilities, see the MHMM Properties section in the Supplementary Material. Once the MHMM was composed, it was then modified to include phylogenetic footprinting capabilities, as described in the following subsection.

2.2 Adding phylogenetic footprinting

Each of the MHMM states described in the previous section corresponds to a column in a multiple-sequence alignment. However, in order to enable SAPF to not only model sequence alignments but also to distinguish between functional and neutral sequence, we double the number of states in the HMM. We refer to the new states as either fast states (corresponding to higher levels of divergence, seen in neutral sequence) or slow states (corresponding to slower divergence, as a consequence of purifying selection in functional elements). The model resulting from this combination of fast and slow hidden states is the SAPF HMM. A sample

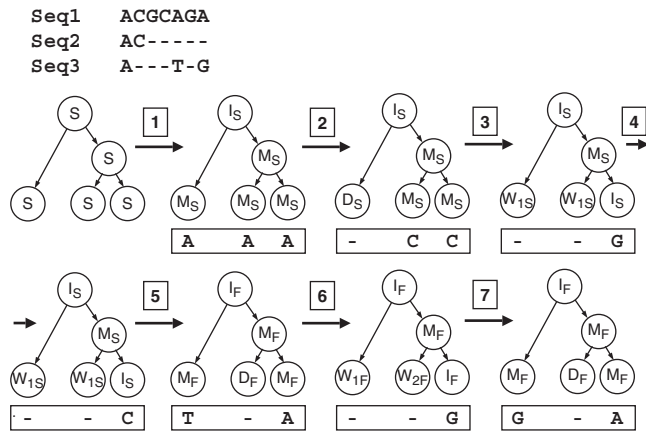


Fig. 2. Sample alignment and a potential representative SAPF HMM state path. Seq 1–3 correspond to the leaves of the tree, ordered **right to left**. Letters in the tree nodes correspond to the state of the branch HMM *leading to* the node (M: Match, I: Insert, D: Delete, W1/W2: Wait1/Wait2, S: Start, E: End). The only exception is the root, which displays the state of the root HMM (see MHMM Properties in the Supplementary Material). State subscripts F and S correspond to fast and slow states. The example features a switch from a slow to a fast state (transition 5), a self-transition in an insertion state (transition 4), and a deletion in sequence two overlapping with an insertion in sequence one (transitions 6–7). The boxes below the states display the letter (or gap) emitted for each leaf sequence.

alignment of three sequences and a potential representative state path through the SAPF HMM are shown in Figure 2.

The annotation of fast or slow must be fixed in all species in an alignment column, and thus the SAPF HMM does not properly model the gain or loss of functional sequence in a single sequence or partial group of sequences. Removing this restriction results in an explosion in the number of hidden states in the HMM, making the algorithm computationally infeasible.

The difference between the fast and the slow states lies in the different branch HMMs used to compose them. While both the fast and slow branch HMM have the same topology, they have different transition and emission probabilities. The probabilities in the slow state (for example, the probability of entering an insert state) are altered to reduce the number of expected evolutionary events in functional sequence. See the Parameters section in the Supplementary Material for a complete discussion of the model parameters used to generate the transition and emission probabilities.

Adding the capability for phylogenetic footprinting is the last step in the creation of the SAPF HMM. The following subsection describes how the SAPF HMM is used to predict the locations of functional elements based on a probability-weighted distribution of alignments.

2.3 Predicting functional elements

SAPF uses standard HMM algorithms to annotate functional elements. Suppose there are m sequences to be analyzed. The forward and backward algorithms (Durbin et al., 1998) are used to analyze the input sequences and to calculate a probability distribution of alignments. This distribution defines, for each group of m nucleotides (one from each of the m species), the posterior probability that they are homologous and therefore should be aligned in a single column. The algorithms also calculate the probability that any alignment column was generated by either a fast state or a slow state. Since laboratory experiments are usually only available for one reference species in a closely related group

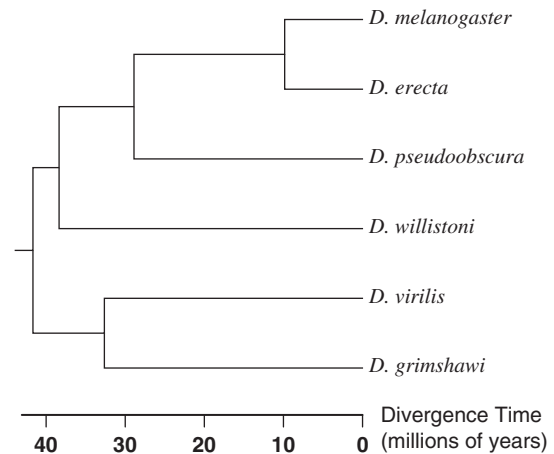


Fig. 3. Phylogenetic tree displaying evolutionary relationships between the *Drosophila* species analyzed by SAPF.

(for example, the *Drosophila melanogaster* genome is the reference for all *Drosophila* species), we have chosen to collapse our results onto one axis and report posterior probabilities for one species, as in (Wasserman et al., 2000). This is accomplished by grouping together all alignment columns containing the same nucleotide in the reference species, and summing over the group to calculate the overall probability that the reference nucleotide was generated from a slow state. We increased the speed of the algorithm by restricting the region of the dynamic programming matrices searched by the forward and backward algorithms (see Corner Cutting, in the Supplementary Material).

We ran SAPF to predict functional elements in both *Drosophila* whole-genome sequences and simulated datasets. The methods used to create these test datasets are discussed in the subsections below.

2.4 Drosophila datasets

In 2007, the *Drosophila* 12 genome consortium finished the sequencing of 10 *Drosophila* species genomes (*Drosophila* 12 Genomes Consortium, 2007; Stark et al., 2007), bringing the number of whole-genome sequences publicly available for analysis to 12. The sequences exhibit a large range of evolutionary distances, as shown in the phylogeny displayed in Figure 3, which is adapted from the ‘accepted *Drosophila* phylogeny’ (Eisen and Holmes, 2008). Some pairs of sequences, such as *D.melanogaster* and *D.simulans* (not pictured), are very closely related, while the evolutionary distance separating *D.melanogaster* and *D.grimshawi* is greater than the evolutionary distance separating any two mammals, when generation time is taken into account (Stark et al., 2007). The diversity in evolutionary distance makes this dataset ideal for phylogenetic footprinting tests, even though SAPF is only currently capable of analyzing data from four species simultaneously. When choosing groups of species to analyze, we attempted to select sequences from different sections of the tree to take advantage of the range of evolutionary diversity in the full dataset. However, we screened out species with promoter regions in which more than a third of previously annotated transcription factor binding sites (TFBS) in *D.melanogaster* (described below) lacked orthologous sites based on a pre-computed set of whole-genome alignments (Dewey et al., 2006), as these sequence may have been too divergent to be informative for phylogenetic footprinting.

An additional advantage to using the *Drosophila* dataset lies in the immense amount of genetics research conducted on *D.melanogaster*. The first assembly for the genome of the ‘fruit fly’ was completed in 2000 (Adams et al., 2000), and since then there has been significant annotation of the genome sequence for known locations of TFBS and *cis*-regulatory

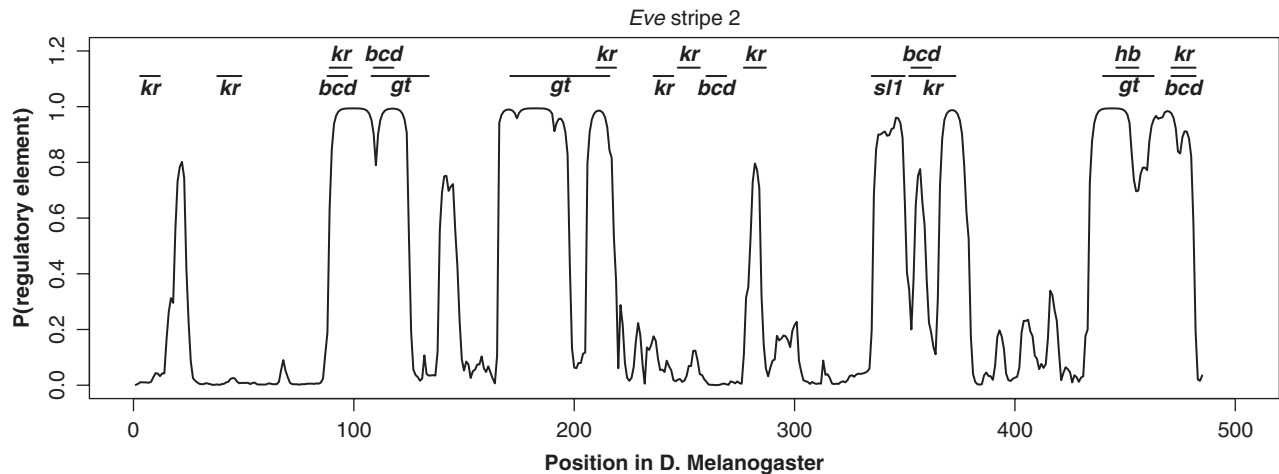


Fig. 4. SAPF predictions and annotated binding sites for *eve* stripe 2 enhancer. For each nucleotide in the *D.melanogaster* sequence, SAPF outputs the probability that the nucleotide was generated by a functional (slow) state. The binding sites in *D.melanogaster* for the transcription factors, bicoid (bcd), hunchback (hb), kruppel (kr), giant (gt), and sloppy-paired 1 (sl1) are shown above the sequence.

modules (CRMs), regions of the genome containing clusters of TFBS. In particular, we relied heavily on two datasets. The REDFly database provided the sequence coordinates of biologically verified CRMs in the *D.melanogaster* sequence (Gallo *et al.*, 2006). Additionally, the FlyReg database curated the results of 201 primary references to annotate the locations of 1367 TFBS in the *D.melanogaster* sequence (Bergman *et al.*, 2005).

The homeodomain encoding *eve* protein is crucial in early embryonic development in *Drosophila*, where it plays a key role in establishing early segmentation in the embryo (Ludwig, 1998). The developing embryonic blastoderm exhibits *eve* expression in a pattern of seven transverse stripes. Though the transcriptional machinery governing this expression pattern is complex, the locations of the enhancer regions governing the expression of each stripe have been experimentally determined, and in many cases, the exact locations of the TFBS have been annotated. We chose to first test SAPF on this region because of the existing high quality laboratory annotations that can be used to test our predictions.

The enhancer controlling the expression of the second stripe is one of the best characterized transcriptional enhancers in *Drosophila*, containing 19 laboratory-annotated binding sites in FlyReg, and is located between approximately -1.5 and -1.0 kb upstream of the transcriptional start site of the *eve* gene in *D.melanogaster* (Bergman *et al.*, 2005; Ludwig, 1998). We ran SAPF on this genomic region using sequence from *D.melanogaster*, *D.erecta*, *D.pseudoobscura* and *D.willistoni*. The exact sequence coordinates for *D.melanogaster* were obtained from the REDFly database, and the sequences for the other species were extracted from a set of pre-computed whole-genome alignments (Dewey *et al.*, 2006).

2.5 Simulated datasets

To further test the model and to verify our results on *Drosophila* sequence data, we created two simulated datasets. Each set consisted of 10 alignments, each 1000 columns long. Each alignment was simulated by sampling a random path through the SAPF HMM, a process which returned both an alignment and annotation for each base as either functional or neutral sequence. SAPF assumes a geometric distribution on the lengths of functional and neutral elements, and the sample alignments had a small proportion of functional elements with a length of < 5 bp. Since binding sites of length < 5 bp are biologically unrealistic, these functional elements were removed from the simulated alignments. The first dataset consisted of alignments sampled using the assumed

species tree and estimated parameter values from the *eve* stripe 2 dataset. The second dataset consisted of alignments with exactly the same species tree, but with modified parameter values. In order to increase the uncertainty in the alignment of functional elements, the difference between the functional and neutral value for three parameters was halved (see Parameters in the Supplementary Material for a complete discussion). The resulting parameter values were similar to the estimated parameter values for the *eve* mesodermal enhancer *Drosophila* sequence dataset. The consequence of this parameter change was a significant increase in the number of evolutionary events in the functional regions, and a resulting increase in the alignment uncertainty in these regions. One advantage of using simulated data as opposed to actual sequence data is that the annotation of functional and neutral sequence is known exactly from the simulation.

3 RESULTS

We ran SAPF on both simulated datasets and recently sequenced *Drosophila* genome sequences in order to test the hypothesis that summing over a distribution of alignments, as opposed to analyzing a single alignment, can improve the quality of functional element predictions.

3.1 *Drosophila* sequence data

SAPF outputs a probability, for each base in *D.melanogaster*, that the base was generated by a functional element state in the SAPF HMM. These probabilities are plotted in Figure 4, along with the experimentally verified locations for the 19 binding sites.

The results indicate that for many of the binding sites, SAPF correctly annotated the bases as functional elements with high probabilities. In this specific dataset, many of the binding sites are extremely well conserved, and thus SAPF is able to make highly certain predictions: 12 binding sites contain bases assigned a posterior functional probability of $> 95\%$, and two others contained bases with posterior probabilities of $> 80\%$. While the five remaining binding sites were incorrectly annotated as neutral, it is likely that functional orthologs do not exist

in all species. Four were initially characterized as ‘low-affinity’ Kruppel binding sites (bases 3–12, 38–49, 236–245, 247–257) (Stanojevic *et al.*, 1991), which may indicate reduced functionality. Additionally, one bicoid binding site (bases 260–269) was postulated to be recently evolved in *D.melanogaster* due to an absence of orthologous sequence in both closely and distantly related *Drosophila* species (Kreitman and Ludwig, 1996; Ludwig, 1998). None of the five binding sites have orthologous sequences in either *D.pseudoobscura* or *D.willistoni*, and thus cannot be identified by phylogenetic footprinting approaches.

SAPF predicts two binding sites that were not previously annotated as functional regions (bases 19–24, 139–147). Both sites are well-conserved in all four sequences, adjacent to previously identified binding sites, and are thus interesting candidates for further experimental study.

For each SAPF run, we examined the area under a receiver operating characteristic (ROC) curve area under the curve (AUC) in order to quantify the predictive accuracy of our results while accounting for both sensitivity and specificity. The methodology used to produce the curves is discussed in the ROC Curves section in the Supplementary Material. The AUC has a maximum value of 100%, while a value of 50% implies that the predictive ability is no better than random guessing. This statistic enables us to examine the benefit of summing over a distribution of alignments. For each dataset, we first ran SAPF to predict functional elements, estimate parameters, and construct a single alignment that summarized the distribution, which we refer to as the maximum posterior product (MPP) alignment. This alignment was defined as the set of alignment columns composing a full alignment that resulted in the maximum product of posterior alignment probabilities. The MPP alignment is a better summary of the alignment distribution than the standard Viterbi alignment since it is computed directly from the posterior probabilities of the distribution, and has also been found to show fewer alignment biases (Lunter *et al.*, 2007). We then ran a restricted version of SAPF in which only the MPP alignment was considered, using the previously estimated parameter set. By constructing ROC curves for the two sets of results, we can assess the difference in predictive ability.

Figure 5a displays these two ROC curves for the *eve* stripe 2 CRM. The two curves, and their corresponding AUC values, are almost identical, indicating that summing over a distribution of alignments did not significantly improve accuracy. However, this can be explained by the lack of uncertainty in the alignment of the TFBS regions. As discussed above, many of the TFBS are exceptionally well conserved in all species, thus each of these regions has a single correct alignment with little uncertainty. The incorrectly annotated binding sites do not have identifiable orthologous sequence in at least two of the four sequences, and thus they are almost impossible to detect using footprinting approaches. In both of these cases, analyzing a distribution of alignments cannot be expected to improve the performance of the predictive tool. Only in cases where there is uncertainty in the alignment would we expect SAPF to exhibit a significant improvement by summing over a distribution of alignments.

We present our results on two additional CRMs that also control regulation of the *eve* gene. The first CRM controls *eve* expression in the third and seventh transverse stripe of the blastoderm expression pattern. To augment the uncertainty

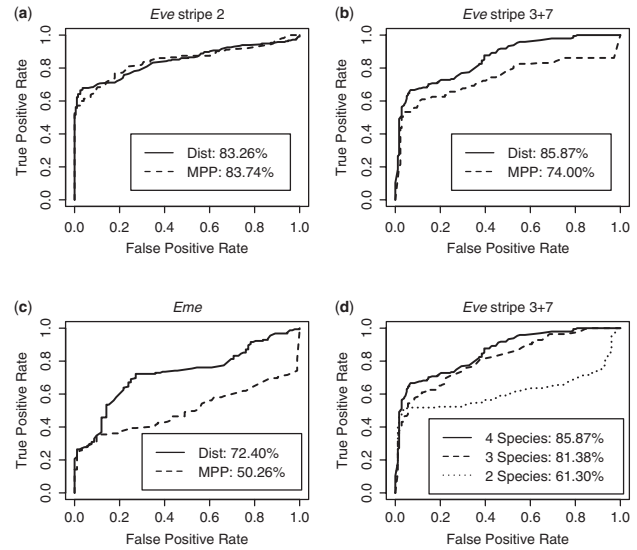


Fig. 5. (a)–(c) ROC curves for predictions created from summing over an alignment distribution, or from analyzing a single MPP alignment for three CRM regions controlling expression of the *eve* gene. (d) ROC curves for predictions created by analyzing different numbers and groupings of *Drosophila* species for the *eve* stripe 3 + 7 region. See ROC Curve Methodology for a complete discussion on how the curves were created. The figure legends show AUC values for all curves.

in the alignment of the functional regions, SAPF was run on sequence data from *D.melanogaster*, *D.erecta*, *D.willistoni* and *D.virilis*. With the exception of the first two, each pair of these sequence diverged more than 36 million years ago. We analyzed this region with the same method used for *eve* stripe 2 enhancer, and the results are shown in Figure 5b. The ROC analysis demonstrates a significant benefit to the accuracy of the functional predictions by summing over a distribution of alignments. Compared to analyzing the single MPP alignment, summing over a distribution increases the AUC statistic by 11.87%.

A binding site for the transcription factor Hunchback, located in the *eve* stripe 3 + 7 CRM, exemplifies the benefit of analyzing multiple alignments when there is alignment uncertainty in regions containing functional elements. Figure 6 displays two potential alignments of the region containing the Hunchback binding site, denoted by a grey box. The MPP alignment, calculated by SAPF, is shown in Figure 6a, but fails to correctly align the binding site in *D.willistoni*. An alternate alignment, shown in Figure 6b, provides a potentially correct alignment of the binding site in all four species. Both alignments are plausible, and the UCSC alignment of 12 *Drosophila* genomes misaligns the *D.willistoni* sequence in the same way as SAPF’s MPP alignment. The existence of two plausible but very different alignments of a binding site is an example of alignment uncertainty in a functional region. When SAPF analyzes only the MPP alignment, it fails to detect the hunchback binding site due to this alignment error. However, when SAPF analyzes an alignment distribution, the alternate alignment in Figure 6b is considered. This enables SAPF to correctly annotate the binding site.

The hypothesis that summing over a distribution of alignments can improve functional element predictions is supported further by the results from the *eve* mesodermal enhancer (*eme*)

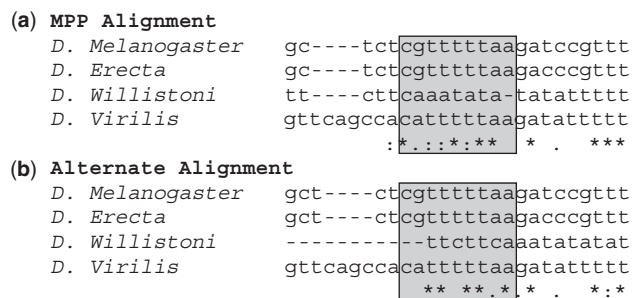
Hunchback binding site, *eve* stripe 3+7 CRM

Fig. 6. Two possible alignments of a region in the *eve* stripe 3+7 CRM. While both alignments are plausible, only (b) aligns the binding site in all four species. Posterior probabilities of the most likely alignment column range from 43%–72% in the binding site region, exhibiting the uncertainty in the alignment. This example demonstrates the necessity for analyzing multiple alignments in order to predict functional elements.

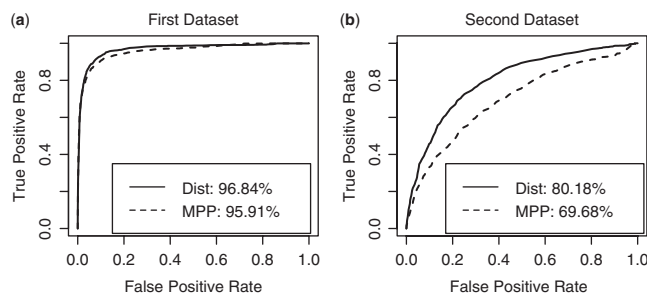


Fig. 7. ROC curves for predictions created from summing over an alignment distribution, or from analyzing a single MPP alignment for two simulated datasets. (a) Dataset was created with the estimated parameters from the *eve* stripe 2 dataset. (b) Dataset was created with modified parameters to increase uncertainty in the alignment of functional regions.

CRM. SAPF was run on on sequence data from *D.melanogaster*, *D.erecta*, *D.willistoni* and *D.grimshawi*. As in the *eve* stripe 3+7 dataset, these sequences are very distantly related with the exception of the first two. The level of conservation among many of the 13 annotated binding sites is extremely poor, and is thus accompanied by high levels of uncertainty throughout all regions of the alignment. Figure 5c reveals that analyzing the distribution of alignments significantly increased the accuracy of the predictions, augmenting the AUC statistic by 22.14%. The dangers of using a single alignment to make functional predictors are apparent when observing that for the MPP alignment predictions, the AUC value of 50.26% is virtually identical to that of a random predictor.

3.2 Simulated sequence data

The results of the ROC analysis for the first simulated dataset are shown below in Figure 7a. The results exhibit two important points. The first is that the quality of the predictions is extremely good, with an AUC value of over 95%. Secondly, the AUC is nearly identical for predictions made from a

distribution of alignments, and predictions based off the MPP alignment. This is consistent with the results obtained from the *eve* stripe 2 dataset.

In the second simulated dataset, the parameter values for the functional sequence states have been altered to create significantly more evolutionary events (substitutions, insertions and deletions) in the functional elements. Due to the increased uncertainty in the alignment of these regions, we expected that SAPF would perform significantly better by summing over a distribution of alignments. We also expected that the increased number of mutations in the functional elements should make them more difficult to detect. The ROC analysis of the second dataset, shown in Figure 7b, confirms both of these predictions. Altering the parameter values reduces the AUC value for the distribution-based predictions curve to slightly over 80%. Additionally, SAPF performs significantly better when summing over a distribution of alignments than it does when analyzing the single MPP alignment alone. The increase in AUC by 10.50% demonstrates SAPF's potential benefit when there is uncertainty in the alignment of the functional regions.

3.3 The effect of multiple sequences

To demonstrate the additional potential benefit of adding multiple sequences, we ran SAPF on different numbers of species while analyzing the *eve* stripe 3 + 7 locus. In addition to the previously mentioned analysis with four species (*D.melanogaster*, *D.erecta*, *D.willistoni* and *D.virilis*), we also ran the analysis using only three species (excluding *D.virilis*), and ran one final analysis including just two species (*D.melanogaster* and *D.virilis*). The ROC analysis shown in Figure 5d demonstrates the additional benefit of adding more species to the analysis, in particular, the significant jump in predictive ability gained from going from two species to three. All curves were generated from predictions based on a distribution of alignments. While we expect that the benefit from adding additional species information may be highly variable, these results imply that significant amounts of phylogenetic information could be lost by focusing on only a small number of sequences.

4 CONCLUSION

Our results have demonstrated, for both simulated and *Drosophila* sequence data, the potential of combining statistical alignment with phylogenetic footprinting to improve the accuracy of regulatory signal detection. We have shown that this benefit increases as the uncertainty in the alignment of functional regions increases, and we have also demonstrated the benefit of analyzing multiple sequences as opposed to a pairwise alignment.

Improvements to SAPF can be made, especially relating to the algorithm's speed and efficiency. Multiple alignment is a *np*-hard problem, and adding phylogenetic footprinting techniques slows the algorithm further. Due to the large number of states in the SAPF HMM, it is currently only possible to analyze data from four different species. While this was sufficient to demonstrate the potential improvement of summing over a distribution of alignments, it places SAPF at a disadvantage to other methods

such as phastCons that can analyze data from significantly greater numbers of species.

One approach that has been popular in recent statistical alignment studies has been to use Markov Chain Monte Carlo simulation techniques to approximate the alignment probability distribution. Creating a Gibbs sampler that sequentially sampled a path through the SAPF HMM (representing both an alignment and annotation) and parameter values may enable us to approximate distributions for these random variables while analyzing significantly greater numbers of species. We hope this technique will enable us to analyze data from the full 12-species *Drosophila* species tree.

Improving the efficiency of the algorithm may also allow us to remove the requirement that the fast/slow annotation must be the same for all species in a given alignment column. Loosening this restriction would allow SAPF to appropriately model gains and losses of binding sites in different species, which would not only increase the quality of the functional element predictions, but also could shed further light on the evolution of transcription factor binding sites in the *Drosophila* genus.

ACKNOWLEDGEMENTS

We thank István Miklós, Rune Lyngsø and Gerton Lunter for helpful discussion. R.S. is funded by the Rhodes Trust, UK.

Conflict of Interest: none declared.

REFERENCES

- Adams,M. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185.
- Bergman,C. *et al.* (2005) *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.
- Boffelli,D. *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
- Bradley,R. and Holmes,I. (2007) Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics*, **23**, 3258–3262.
- Cliften,P. *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
- Dewey,C. *et al.* (2006) Parametric alignment of *Drosophila* genomes. *PLoS Computat. Biol.*, **2**, e73.
- Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.
- Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, New York.
- Eisen,M. and Holmes,I. (2008) *Phylogeny*, <http://rana.lbl.gov/drosophila/wiki/index.php/Phylogeny> (last accessed date January 15 2008).
- Gallo,S. *et al.* (2006) REDfly: a Regulatory element database for *Drosophila*. *Bioinformatics*, **22**, 381–383.
- Holmes,I. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, **19**, 147–157.
- Holmes,I. (2005) Using evolutionary expectation maximization to estimate indel rates. *Bioinformatics*, **21**, 2294–2300.
- Holmes,I. (2007) Phylocomposer and phylodirector: analysis and visualization of transducer indel models. *Bioinformatics*, **23**, 3263–3264.
- Holmes,I. and Rubin,G. (2002) An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.*, **317**, 753–764.
- Kreitman,M. and Ludwig,M. (1996) Tempo and mode of even-skipped stripe 2 enhancer evolution in *Drosophila*. *Sem. Cell Dev. Biol.*, **7**, 583–592.
- Ludwig,M. (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development*, **125**, 949–958.
- Lunter,G. *et al.* (2007) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.
- Lunter,G. (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, **23**, i289.
- Miklós,I. *et al.* (2004) A “Long Indel” Model For Evolutionary Sequence Alignment. *Mol. Biol. Evol.*, **21**, 529–540.
- Pollard,D. *et al.* (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**, 376.
- Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034.
- Sinha,S. and He,X. (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput. Biol.*, **3**, e216, doi:10.1371/journal.pcbi.0030216.
- Stanojevic,D. *et al.* (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science*, **254**, 1385–1387.
- Stark,A. *et al.* (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, **450**, 219–232.
- Tagle,D. *et al.* (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
- Thorne,J. *et al.* (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, **33**, 114–124.
- Thorne,J. *et al.* (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, **34**, 3–16.
- Wang,J. *et al.* (2006) MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics*, **7**, 292.
- Wasserman,W. *et al.* (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, 225–228.
- Zhu,J. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.