BIOINFORMATICS

Vol. 23 ECCB 2006, pages e44–e49
doi:10.1093/bioinformatics/btl305

Simultaneous alignment and annotation of *cis*-regulatory regions

Abha Singh Bais*, Steffen Grossmann and Martin Vingron

Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195, Berlin, Germany

ABSTRACT

Motivation: Current methods that annotate conserved transcription factor binding sites in an alignment of two regulatory regions perform the alignment and annotation step separately and combine the results in the end. If the site descriptions are weak or the sequence similarity is low, the local gap structure of the alignment poses a problem in detecting the conserved sites. It is therefore desirable to have an approach that is able to simultaneously consider the alignment as well as possibly matching site locations.

Results: With SimAnn we have developed a tool that serves exactly this purpose. By combining the annotation step and the alignment of the two sequences into one algorithm, it detects conserved sites more clearly. It has the additional advantage that all parameters are calculated based on statistical considerations. This allows for its successful application with any binding site model of interest. We present the algorithm and the approach for parameter selection and compare its performance with that of other, non-simultaneous methods on both simulated and real data.

Availability: A command-line based C++ implementation of SimAnn is available from the authors upon request. In addition, we provide Perl scripts for calculating the input parameters based on statistical considerations.

Contact: bais@molgen.mpg.de

1 INTRODUCTION

Using cross-species comparisons for annotating *cis*-regulatory regions is a well-established approach in computational genomics. It is based on the rationale that functionally relevant sequence features evolve slower than non-functional ones. A method implementing this should be able to align orthologous promoter or enhancer sequences and simultaneously predict conserved transcription factor binding sites (TFBSs). Although many methods have been proposed that provide a combination of conservation and TFBS annotation (for reviews see Ureta-Vidal *et al*., 2003; Wasserman *et al*., 2004), to our knowledge none of them achieves this simultaneously.

In general, existing methods solve the problem in two main steps. In one step, conserved regions between two orthologous sequences are extracted using a method-specific alignment algorithm and a conservation criterion. In a separate step, log-likelihood based models called position-specific scoring matrices (PSSMs) are used to scan individual sequences for putative TFBSs. Finally, the alignment and annotation results are combined to predict conserved TFBSs. Representative examples include ConSite (Sandelin *et al*., 2004) and CisOrtho (Bigelow *et al*., 2004). Some methods extend this general strategy by incorporating additional information. This can either be gene expression data like in oPossum (Ho Sui *et al*., 2005), clustering of TFBSs in conserved regions as in rVista (Loots *et al*., 2004) or relative positional preferences, Footer (Corcoran *et al*., 2005).

Another class of methods uses prior knowledge of TFBSs to construct the alignments. Putative TFBS hits on the single sequences are paired and used as anchors for producing either global (Berezikov *et al*., 2004) or local alignments (Michael *et al*., 2005). While ConReal focuses on generating an ordered chain of conserved TFBSs, thus not aligning regions that do not contain them, Siteblast is a BLAST-like heuristic where the TFBS hits are used as seeds. The method of Hallikas *et al*. (2006) also falls in this category. Here, the sequence of hit pairs is aligned using a scoring scheme that considers clustering of sites, binding affinity and conservation, though the underlying sequences themselves are not aligned.

Other approaches like Monkey (Moses *et al*., 2004) explicitly take into account evolutionary properties of the TFBSs, but still perform the alignment independent of the annotation step. Recently, *ab initio* methods have also been developed which use an available alignment and evolutionary constraints on the binding sites (Sinha *et al*., 2004; Siddharthan *et al*., 2005).

In summary, most of the methods depend on a predetermined optimal alignment for deciding whether a hit pair is predicted as conserved or not. If the optimal alignment fails in detecting such a conserved hit pair, slight local modifications in the alignment might suffice to remedy this situation. Hence it is desirable to have a method that suitably combines the alignment and annotation steps to allow for this flexibility. We propose an extended pairwise alignment algorithm that provides this direct combination of the two steps. It introduces the possibility of annotating parts of an alignment as paired profiles and extends the scoring scheme appropriately. Since we have a statistically motivated approach to determine the additional scoring parameters, the calculation of the optimal alignment in this extended model allows for local rearrangements in the alignment to make conserved TFBSs stick out. The algorithm is implemented as the SimAnn program and is available on request.

The rest of the article is organized as follows. In Section 2 we describe in detail our extended alignment model with the necessary algorithmic modifications. Our theory for deriving profile related scoring parameters follows. At the end of the section we describe the strategy for a systematic validation of our approach. The results of this validation are given at the beginning of Section 3. Finally, we present a case study analyzing the *Drosophila* even-skipped stripe 2 enhancer region. We discuss the applicability and potential of our method at the end of the article.

2 METHODS

2.1 Extended Alignments

The aim of our method is to combine a locally optimal alignment of two sequences with an annotation with evolutionarily conserved pairs of profiles.

*To whom correspondence should be addressed.

We therefore add the possibility of assigning parts of the alignment directly to such perfectly aligned pair-profiles. This extension in the alignment scheme is introduced to allow for a different scoring of these pair-profiles as follows.

Assume that we wish to search for conserved instances of a profile P of length l. A stretch of l consecutive gaplessly aligned positions can be scored in the extended alignment model in two possible ways: either by scoring each aligned pair with the standard substitution scoring matrix S, or by using a profile scoring array (PSA) and subtracting a profile penalty p. The PSA assigns a score to every pair of strings of length l and reflects how well the gapless alignment of this pair fits to the motif described by P. The profile penalty is a tuning parameter meant to maintain the balance between the two alternatives.

Figure 1 illustrates the extended alignment approach through an example of two sequences x and y. In Figure 1(a) a possible standard alignment is depicted wherein the putative hit of a profile (here simply the TATA-box) is not predicted as a conserved hit. In Figure 1(b) the situation changes, since now the alignment can shift gaps to bring forth a putative conserved hit.

The Smith–Waterman algorithm for the determination of optimal local alignments is modified in a straightforward way to incorporate the additional pair-profile states. In the case of linear gap penalties and a single profile, the modified recursion rule is

$$\mathtt{M}(i,j) = \max\begin{cases} 0, \\ \mathtt{M}(i-1,\ j-1) + \mathtt{s}(x_i, y_j), \\ \mathtt{M}(i-1,\ j) - g, \\ \mathtt{M}(i,\ j-1) - g, \\ \mathtt{M}(i-l+1,\ j-l+1) - p \\ \quad + \mathrm{PSA}((x_{i-l+1}\cdots x_i), (y_{j-l+1}\cdots y_j)) \end{cases}$$

where g is the gap penalty. Extensions to multiple profiles and affine-linear gap case is equally straightforward.

2.1.1 Calculation of scoring parameters We now describe our derivation of the profile related scoring parameters PSA and p in more detail. We assume that the substitution scoring matrix S is given in the form of a log-likelihood ratio of a distribution q for evolutionarily related letter pairs with respect to an independent sampling of two letters from a background distribution π. In the same manner, the profile scoring array PSA is defined as the log-likelihood ratio of a distribution on pairs $(\mathbf{u}, \mathbf{v})$ of strings of length l with respect to the same background distribution π. Here, the distribution on the string pairs should reflect the properties of the profile. Hence, we start with the position-specific letter distribution $\mathrm{P} = (\mathrm{P}^1, \ldots, \mathrm{P}^l)$ of the profile and consider two strings $\mathbf{u}$ and $\mathbf{v}$ to be sampled independently from P. In the corresponding background distribution all letters occurring in the strings are sampled independently from π. More formally, this leads to

$$\begin{aligned} \mathrm{PSA}(\mathbf{u},\mathbf{v}) &:= \sum_{i=1}^{l} \log\left(\frac{\mathrm{P}^i(u_i)\mathrm{P}^i(v_i)}{\pi(u_i)\pi(v_i)}\right) \\ &= \sum_{i=1}^{l} \log\left(\frac{\mathrm{P}^i(u_i)}{\pi(u_i)}\right) + \sum_{i=1}^{l} \log\left(\frac{\mathrm{P}^i(v_i)}{\pi(v_i)}\right) \\ &=: \mathrm{PSSM}(\mathbf{u}) + \mathrm{PSSM}(\mathbf{v}), \end{aligned} \tag{1}$$

where PSSM denotes the position-specific scoring matrix.

The additive form comes from the fact that we sampled the two strings independently in the pair model chosen. Other approaches, for example considering a pair of evolutionarily related samples from P, can also be used but are not studied in this article.

Recall that the profile penalty p has been introduced for a fine-tuning of the balance between the two gapless scoring alternatives of the two strings $\mathbf{u}$ and $\mathbf{v}$. Whenever $\mathrm{PSA}(\mathbf{u},\mathbf{v}) - p > \sum_{i=1}^{l} \mathtt{s}(u_i, v_i)$, the corresponding stretch in the alignment is assigned to the profile rather than to l successive substitutions. After rewriting, we see that this is equivalent to

$$\mathrm{LLR}_{\mathrm{P}^2,\, q^l}(\mathbf{u},\mathbf{v}) := \log\frac{\mathrm{P}(\mathbf{u})\mathrm{P}(\mathbf{v})}{\prod_{i=1}^{l} q(u_i, v_i)} > p. \tag{2}$$

Since the calculation of all scores involved is based on the same background model π it cancels out here. The penalty p can be interpreted as a cutoff in a log-likelihood ratio test. Now the log-likelihood ratio directly compares the pair profile measure P^2 and the measure q^l which arises from independently sampling l evolutionarily related letter pairs from q.

Techniques similar to the ones described in Rahmann *et al*. (2003) allow us now to calculate the exact distribution of $\mathrm{LLR}_{\mathrm{P}^2,\, q^l}(\mathbf{u},\mathbf{v})$ under the two measures P^2 and q^l and therefore to make a statistically justified choice of p.

In the following, we use three natural choices. First, we choose p such that for a pre-specified level α the type-I error probability $\mathbb{P}_{q^l}(\mathrm{LLR}_{\mathrm{P}^2,\, q^l}(\mathbf{u},\mathbf{v}) > p)$ is smaller than α. We call this the *level* α type-I error penalty. Second, we choose p such that the corresponding type-II error probability $\mathbb{P}_{\mathrm{P}^2}(\mathrm{LLR}_{\mathrm{P}^2,\, q^l}(\mathbf{u},\mathbf{v}) < p)$ is smaller than a pre-specified level β. We call this the level β type-II error penalty. Finally, we choose p such that the two error probabilities are equal, in which case we speak of the balanced penalty. We refer the reader again to the work of Rahmann *et al*, (2003) for details on the justification of these choices and the algorithmic techniques needed to calculate the exact error probabilities.

2.1.2 Implementation SimAnn is a command-line based C++ implementation of the extended alignment algorithm. It can handle multiple profiles and affine-linear gap-penalties and is available from the authors upon request. It comes with a set of Perl scripts providing functionality for the calculation of the profile related scoring parameters.

2.2 Simulation setting

We give an initial validation of our approach in the following simulation setting. We generate a large set of evolutionarily related sequence pairs into each of which we implant a pair of motifs sampled from a fixed profile P. The correct alignments and positions of the implanted motifs are stored for later evaluation. The raw sequence pairs are analyzed with SimAnn and two multi-step approaches to detect conserved binding sites. All methods are provided with the profile P from which the implanted motifs have been sampled. For each method there is a single parameter which balances its sensitivity and specificity. We vary this parameter in order to determine the respective receiver operator characteristics (ROC). For SimAnn this parameter is the profile penalty p. Hence, we can also use the ROC curves to assess the quality of our theoretically determined profile penalty choices. This analysis is carried out for different values of sequence relatedness and different quality of the implanted profiles.

2.2.1 Construction of simulated data For a fixed evolutionary distance and a fixed profile we adopt the following strategy to generate a set of sequence pairs.

We use the software program Rose version 1.3 (Stoye *et al*., 1998) to simulate sequence pairs at specified evolutionary distance (called relatedness in Rose) together with their true alignments. The sequences are specified to be at the leaves of a simple depth one binary tree with branch lengths proportional to the distance. We set the indel threshold to 0.002 for a better balance between substitutions and insertions/deletions than with the default value. All other parameters are set to the default DNA settings. The final set consists of 50 sampled sequence pairs of an average length of 500.

The profile, given as a position-specific count matrix, is first converted to a regularized position-specific probability matrix (PSPM) as in Rahmann *et al*. (2003). For each sequence pair two motifs are sampled independently from this PSPM. The true alignment of the sequence pair is cut at a random position and one of the sampled profiles is inserted into each sequence.

We repeat this construction for relatedness values ranging from 10 to 50 at steps of 10 and for three profiles of differing quality, resulting in a total of 15 different datasets. As a measure of profile quality we use the balanced quality as described in Rahmann *et al*. (2003) where the type-I and type-II errors are equal. The matrices, taken from the TRANSFAC (Matys *et al*., 2003) database, are M00395 (poor quality, 0.199), M00690 (medium quality, 0.622) and M00360 (good quality, 0.967).

Fig. 1. Two possible alignments of sequences *x* and *y*. In the standard alignment model the optimal alignment might look as in (**a**), where S, D and I represent substitutions, deletions and insertions respectively. The additional annotation options in the extended alignment model can shift the gaps locally to better highlight the conserved *cis*-regulatory elements as in (**b**).

2.2.2 Multi-step approaches We have implemented two multi-step approaches to detect conserved binding sites. Both methods first align the two sequences using the standard Smith–Waterman algorithm (Smith *et al*., 1981) with affine gap penalties. For each profile specified, both sequences are scanned for putative hits using the scheme described by Rahmann *et al*. (2003). Here, the choice of the score cutoff influences the number of accepted hits and can be used as a parameter to control the final balance between sensitivity and specificity. The hits are then mapped onto the alignment as a basis for filtering out the conserved hit pairs. The two approaches differ only with respect to this filtering. We distinguish between a Relaxed and a Strict filtering.

Relaxed filter. A hit pair is marked as conserved if the mapped hit on the first sequence overlaps positively with that on the second sequence in the alignment, irrespective of the number of gaps in the mapped regions of the alignment.
Strict filter. A pair is marked as conserved only if the mapped hits contain no gaps, and the hit on the first sequence is perfectly aligned with that on the second sequence.

By considering both the Relaxed and the Strict filters, we cover two extremes of the spectrum. While the Relaxed filter provides an over-estimate by allowing unlimited number of gaps in the aligned hits, the Strict filter provides a lower estimate with no leniency for alignment errors.

2.2.3 Parameter choice We can use the same parameters for the standard alignment part of SimAnn and the Smith–Waterman alignment algorithm underlying the two multi-step approaches. This ensures that the differences observed in the comparison of the three approaches can directly be attributed to those aspects of the methods which are added onto the basic alignment part. In the case of SimAnn this is the introduction of the pair profile states into the alignment algorithm and their special scoring.

To get the correct standard alignment parameters first the substitution matrix that fits to the chosen evolutionary distance is determined. This derivation is straightforward because Rose uses a Jukes–Cantor substitution model and a uniform background letter distribution. To find the appropriate affine gap penalty scheme, we first restrict ourselves to the set where the gap extension penalty is 1/10 of the gap open penalty. A set of sequence pairs at the fixed evolutionary distance is generated with Rose as described above. All generated sequence pairs are realigned under different gap open penalties and the proportion of gaps in the true and the recomputed alignments is compared to determine the optimal gap open penalty.

3 RESULTS

3.1 Simulated data

We analyze each of the generated sequence sets with SimAnn and the multi-step approaches in terms of ROC curves. These are obtained by varying the profile penalty in SimAnn and the PSSM score cutoff in the multi-step approaches over a wide range. True and false positive rates (TPR and FPR) are calculated as follows. If the implanted motif is detected as a conserved hit it is counted as a true positive. So there can be at most one true positive in each of the 50 sequence pairs. Since, in contrast to SimAnn, the multi-step approaches can predict overlapping conserved hits, we define the false positive rate as the relative amount of non-profile sequence covered by predicted conserved profile pairs.

The advantage of varying the profile penalty over a wide range is that we can additionally use the ROC curves to validate the performance of our theoretically calculated profile penalties. These penalties corresponding to the three proposed values (level 0.05 type-I error, level 0.05 type-II error and balanced) are highlighted in color on the ROC curves for SimAnn. Results for two evolutionary distances (10 and 40) and the three profiles of different quality are shown in Figure 2.

With increasing evolutionary distance and decreasing profile quality it gets more difficult to detect the implanted motifs, and all the three methods reflect this. Moreover, since both multi-step approaches are based on the same alignment and annotation results, the Relaxed filter (blue) performs much better than the Strict filter (green), as can be seen from the ROC curves.

It is striking that the true positive rate for SimAnn (black) decreases at extremely low profile penalties. This is understandable since SimAnn cannot predict overlapping instances of conserved pairs as opposed to the multi-step approaches. At low penalties, SimAnn tries to fill the alignment with as many non-overlapping instances of pair profiles as possible, and thereby looses the annotations that it correctly predicted at higher penalties. This underlines the importance of a correct choice of the profile penalty.

As can be seen from the cutoffs highlighted in Figure 2, the balanced profile penalty (red cross) and the type-II error penalty at level 0.05 (magenta cross) both fall into a region where true and false positive rates show reasonable combinations. Thus, a balanced profile penalty can be chosen when high sensitivity is required while the type-II error penalty at level 0.05 can be chosen for needs of high specificity. We therefore have shown that a good choice for all necessary parameters in SimAnn can be made based on theoretical considerations. This avoids the need for *ad hoc* decisions or extensive simulation studies for every incoming profile. It also enables us to use SimAnn such that we achieve the highest performance.

In comparison with the multi-step methods, SimAnn performs comparably well in all situations and has a clear advantage as evolutionary distance increases or profile quality worsens.

It should be stressed here that even though the Relaxed multi-step performs as well as SimAnn, the predicted conserved pairs are not necessarily perfectly aligned in the optimal alignment. They can be interrupted by any number of gaps making them difficult to stand out as a conserved binding site. Contrarily, the predictions from SimAnn are perfectly aligned, gapless pairs of profiles and the conserved binding site is clearly identifiable.

3.2 Extracting conserved binding sites: a case study

As an example we consider the even-skipped stripe 2 region in *Drosophila melanogaster* and *Drosophila pseudoobscura*. This is a well-characterized *cis*-regulatory module (Stanojevic *et al*., 1991;

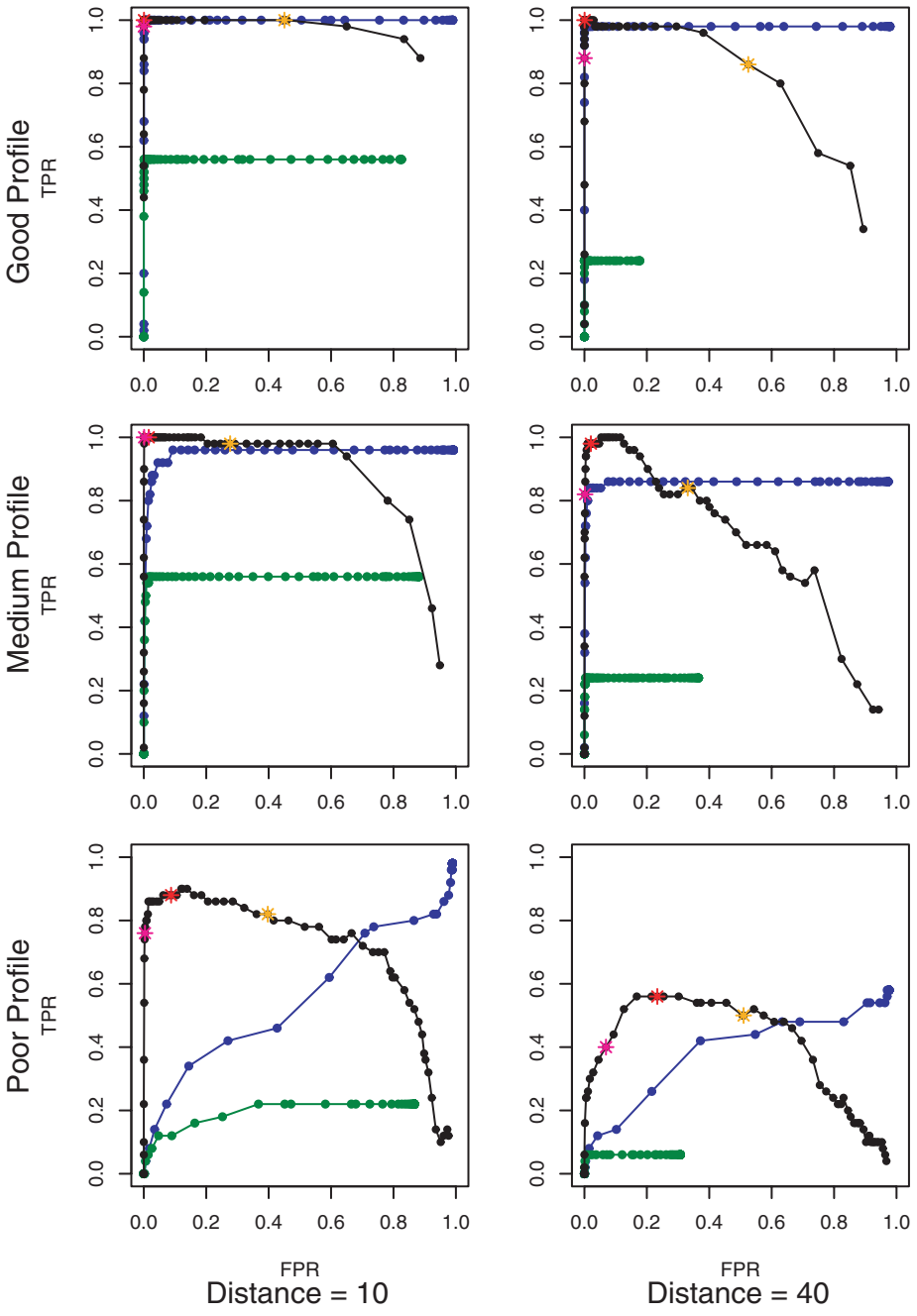

Fig. 2. Performance comparison of SimAnn and the two multi-step approaches at two evolutionary distances and for three profiles of different quality. The ROC curves illustrate the interplay of false and true positive rates on the simulated test sets at varying penalty/cutoff parameters. Color code of the ROC curves: green, multi-step approach/Strict filter; blue, multi-step approach/Relaxed filter; black, SimAnn. On the SimAnn ROC curve the statistically motivated profile penalty choices are highlighted. Color code: orange, type-I error penalty at level 0.05; red, balanced penalty; magenta, type-II error penalty at level 0.05.

Ludwig *et al*., 1998, 2005) containing multiple binding sites for at least four transcription factors: Bicoid, Hunchback, Giant and Krüppel. There are a total of 17 experimentally verified sites for these factors in this region and the corresponding count matrices are available (Rajewsky *et al*., 2002). We retrieved the orthologous enhancer sequences using the Genbank identifiers provided in Ludwig *et al*. (2005). The lengths of the individual sequences amount to 799 in *D.melanogaster* and 1028 in *D.pseudoobscura*.

We compared SimAnn with the multi-step approaches and a third tool called ConSite available online (Sandelin *et al*., 2004). We consider ConSite because it is also a multi-step approach where first alignments are generated and conserved regions extracted. Then, sequences are scanned for putative hits using a score cutoff which does not consider the background letter distribution. Finally, only those hits that are situated in conserved regions and lie at equivalent positions in the alignment are output as conserved pairs.

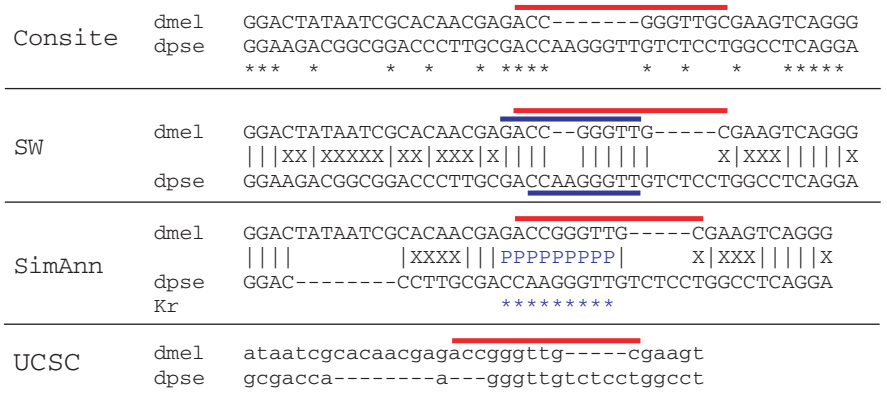

Fig. 3. Alignment region of the Krüppel 4 site. Red indicates the true location, while blue depicts predictions made by the respective methods. SW stands for the Smith–Waterman alignment used for the multi-step approaches.

Both SimAnn and our multi-step approaches are run with the standard HOXD70 substitution scoring matrix with gap open cost of 400 and extension cost of 30. The count matrices describing the relevant factors are preprocessed as described in Section 2.1.1 to calculate the profile related parameters for SimAnn. For sequence scanning within the multi-step approaches, count matrices are converted into scoring matrices and score cutoffs are determined along the lines of Rahmann *et al*. (2003). For ConSite, we use two main parameter settings, the default and with conservation and matrix score cutoff of 70%. Count matrices are same as above. In all methods we count a prediction correct if it overlaps with the known binding site by more than quarter of the length of the PSSM. Overlapping predictions of the same PSSM are counted only once.

Out of the 17 sites, the Relaxed multi-step approach predicts 10 while the Strict predicts 9 sites correctly, with no false positives. The one site that is missed out by the Strict filter owing to gaps in the alignment is the Krüppel 4 site. With ConSite, the default settings yield much fewer predictions, namely 5 out of 17. When the matrix score cutoff is lowered to 70%, this number increases to 10, at the cost of predicting additional 10 false positives.

When SimAnn is run with all four profiles together it predicts 9 sites correctly. We also run SimAnn supplying each profile individually to check whether overlapping binding sites pose a problem and this raised the number of true positives to 11. In both cases we obtained four false positives.

Overall, our multi-step approaches and SimAnn perform very similarly, which is expected since they are based on the same premises algorithmically and parametrically. But ConSite has a slightly poorer performance since the gain in sensitivity by lowering cutoffs results in a drastic increase in the number of false positives, too.

It is worth looking in more detail at the Krüppel 4 site mentioned above because it is the only site which resides in a region of ambiguous alignment. The results of all methods are shown in Figure 3. The Strict multi-step approach and ConSite fail to predict the site because of gaps in the underlying alignment. ConSite predicts it, but only after the matrix score cutoff and the conservation cutoff are reduced to 60% and 40%, respectively. The Relaxed multi-step approach and SimAnn successfully predict the site. However, with SimAnn the nice feature is that the binding site stands out more clearly. Through the UCSC alignment of the site, also shown in Figure 3, one can see that there is no clear correct alignment—the UCSC alignment differs considerably from the rest.

4 CONCLUSIONS

In this article we have introduced a novel integrated approach SimAnn to detect conserved transcription factor binding sites. In SimAnn the alignment and annotation steps are combined in one extended alignment model. This enables the method to locally shift gaps in the alignment to make the conserved hits stick out more clearly. An extended alignment method as SimAnn stands and falls with the choice of the parameters needed in the model. With a statistically founded strategy for parameter selection we have solved this problem in SimAnn and thus can handle any profile of interest.

We demonstrated the applicability of SimAnn via a systematic comparison with other multi-step approaches on simulated data. We showed how SimAnn can predict perfectly aligned conserved hit pairs even in conditions of higher evolutionary distance or poorer profile quality. By analyzing the well-known even-skipped stripe 2 enhancer region in two *Drosophila* species we illustrated the potential of SimAnn in a biological setting.

SimAnn is best suited for detailed analysis of a regulatory region known to be conserved between two species with available information of certain essential transcription factors. Especially when conservation is weak and it is difficult to identify conserved binding sites, SimAnn can assist a lot in understanding the potential regulatory mechanisms in the region. However, for analyzing arbitrarily big conserved regions with a large number of profiles, SimAnn is not particularly suited. The resulting multiple testing problems and the increased complexity of the extended alignment model could hinder performance and well-established multi-step approaches should be preferred.

We are extending the SimAnn approach in various directions. Work is in progress to enable detection of suboptimal alignments along the lines of the Waterman–Eggert algorithm (Waterman *et al*., 1987). Applicability on more real examples is also being evaluated. Our current construction of the profile scoring array, which is based on two independent samples of the profile, is not the only possible approach. A possible alternative would be to introduce evolutionarily related profile samples. With minor changes, the statistical calculations of the profile penalties should work in this case, too.

As a further extension, SimAnn would even be able to use profile scoring arrays that combine binding site descriptions that differ to some extent in the two species aligned. Once available, knowledge about co-evolution of transcription factors and their DNA binding sites could thus be incorporated.

ACKNOWLEDGEMENTS

The authors thank Ho-Ryun Chung for providing us with details of the *Drosophila* example and Stefan Röpcke for careful reading of the manuscript.

REFERENCES

Berezikov,E. *et al*. (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res*., **14**, 170–178.

Bigelow,H.R. *et al*. (2004) CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinfo*., **5**, 27.

Corcoran,D.L. *et al*. (2005) FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res*., **33**, W442–W446.

Hallikas,O. *et al*. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.

Ho Sui,S.J. *et al*. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*., **33**, 3154–3164.

Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*., **32**, W217–W221.

Ludwig,M.Z. *et al*. (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development*, **125**, 949–958.

Ludwig,M.Z. *et al*. (2005) Functional evolution of a *cis*-regulatory module. *PLoS Biol*., **3**, e93.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Michael,M. *et al*. (2005) SITEBLAST—rapid and sensitive local alignment of genomic sequences employing motif anchors. *Bioinformatics*, **21**, 2093–2094.

Moses,A.M. *et al*. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Gen. Biol*., **5**, R98.

Rahmann,S. *et al*. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol*., **2**, article 7.

Rajewsky,N. *et al*. (2002) Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.

Sandelin,A. *et al*. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*., **32**, W249–W252.

Siddharthan,R. *et al*. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol*., **1**, e67.

Sinha,S. *et al*. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Stanojevic,D. *et al*. (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science*, **254**, 1385–1387.

Stoye,J. *et al*. (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

Ureta-Vidal,A. *et al*. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet*., **4**, 251–262.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.

Waterman,M.S. and Eggert,M. (1987) A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *J. Mol. Biol*., **197**, 723–728.