Incorporating evolution of transcription factor binding sites into annotated alignments

Abha S Bais*, Steffen Grossmann and Martin Vingron
Max Planck Institute for Molecular Genetics, Berlin, Germany
**Corresponding author (Fax, 493084131152; Email, bais@molgen.mpg.de)*

Identifying transcription factor binding sites (TFBSs) is essential to elucidate putative regulatory mechanisms. A common strategy is to combine cross-species conservation with single sequence TFBS annotation to yield "conserved TFBSs". Most current methods in this field adopt a multi-step approach that segregates the two aspects. Again, it is widely accepted that the evolutionary dynamics of binding sites differ from those of the surrounding sequence. Hence, it is desirable to have an approach that explicitly takes this factor into account. Although a plethora of approaches have been proposed for the prediction of conserved TFBSs, very few explicitly model TFBS evolutionary properties, while additionally being multi-step. Recently, we introduced a novel approach to simultaneously align and annotate conserved TFBSs in a pair of sequences. Building upon the standard Smith-Waterman algorithm for local alignments, SimAnn introduces additional states for profiles to output extended alignments or annotated alignments. That is, alignments with parts annotated as gaplessly aligned TFBSs (pair-profile hits) are generated. Moreover, the pair-profile related parameters are derived in a sound statistical framework.

In this article, we extend this approach to explicitly incorporate evolution of binding sites in the SimAnn framework. We demonstrate the extension in the theoretical derivations through two position-specific evolutionary models, previously used for modelling TFBS evolution. In a simulated setting, we provide a proof of concept that the approach works given the underlying assumptions, as compared to the original work. Finally, using a real dataset of experimentally verified binding sites in human-mouse sequence pairs, we compare the new approach (eSimAnn) to an existing multi-step tool that also considers TFBS evolution.

Although it is widely accepted that binding sites evolve differently from the surrounding sequences, most comparative TFBS identification methods do not explicitly consider this. Additionally, prediction of conserved binding sites is carried out in a multi-step approach that segregates alignment from TFBS annotation. In this paper, we demonstrate how the simultaneous alignment and annotation approach of SimAnn can be further extended to incorporate TFBS evolutionary relationships. We study how alignments and binding site predictions interplay at varying evolutionary distances and for various profile qualities.

[Bais A S, Grossmann S and Vingron M 2007 Incorporating evolution of transcription factor binding sites into annotated alignments; *J. Biosci*. **32** 841–850]

1. Introduction

A majority of computational approaches that aim to predict transcription factor binding sites employ cross-species comparison to focus on conserved locations. Such a comparison helps in filtering out the vast amount of false predictions arising due to the short and degenerate nature of binding sites. The rationale is that conserved sites are more likely to be biologically relevant.

The standard strategy for combining transcription factor binding sites (TFBS) prediction with conservation is the following. A probabilistic model for binding sites known

Keywords. Alignments; evolutionary models; transcription factor binding sites

Abbreviations used: PSSM, Position-specific scoring matrix; ROC, receiver operator characteristics; SW, Smith-Waterman; TFBS, transcription factor binding sites

as a *profile* [or a position-specific weight matrix (Stormo 2000)] is used to scan individual sequences for putative TFBS hits. These are then combined with an alignment of the two sequences to identify conserved TFBSs. While there are more sophisticated methods that include additional filters like gene expression data or information about clustering of binding site locations (examples include Loots and Ovcharenko 2004; Sui *et al* 2005), in general, methods follow the same underlying multi-step approach. For recent reviews on TFBS prediction methods that use conservation information, *see* Wasserman and Sandelin (2004), Siggia (2005) and MacIsaac and Fraenkel (2006).

Combining single sequence TFBS annotation and sequence alignment usually requires the use of either a conservation criterion (like percentage sequence identity) or/and individual hit score cutoffs (e.g. Lenhard *et al* 2003). Such a setting, which relies on predetermined optimal alignments, may suffer from underlying alignment errors. Especially if the TFBS profile is weak or sequence similarity poor, it is possible that gaps obscure putative conserved TFBSs necessitating local rearrangements in the alignment. In a recent article (Bais *et al* 2007), we introduced a simultaneous alignment and annotation algorithm SimAnn that incorporates profiles into the alignment generation and outputs annotated alignments. Such a simultaneous approach allows for local rearrangements in the alignment to bring forth gaplessly aligned TFBSs.

Again, cross-species comparison to filter conserved TFBSs aims to take advantage of the underlying evolutionary relationship between the sequences. The idea of a conserved pair of TFBSs is that each of the hits should represent the underlying profile and be evolutionarily related. However, simply using the conservation criterion equates to ignoring binding site specific evolutionary characteristics. Similarly, searching separately for high scoring hit-pairs implies assuming independence between the individual hits. While in the former a region that is well-conserved may lead to false predictions arising due to high sequence similarity, in the latter non-consensus nucleotide mutations arising from high sequence divergence might yield low score for a pair, thus missing the correct prediction. It is widely accepted that evolutionary properties of transcription factor binding sites differ significantly from those of the surrounding sequences.

Much research has gone into the study of the evolution of binding sites, (McCue *et al* 2002; Gerland and Hwa 2002; Dermitzakis and Clark 2002; Moses *et al* 2003; Berg *et al* 2004; Ludwig *et al* 2005; Kotelnikova *et al* 2005; Mustonen and Lässig 2005; Wittkopp 2006). It has been shown that (i) binding sites evolve slower than the surrounding sequences and (ii) they exhibit varying rates of evolution at each position. Moses *et al* (2003) use yeast species for analysis and propose the use of an evolutionary model (Halpern and Bruno 1998) for modelling position-specific evolution of TFBSs. Kotelnikova *et al* (2005) conclude from bacterial binding site data that TFBSs exhibit selectional constraints on degenerate positions also. As would be expected, positions with clear bias towards a single nucleotide tend to maintain their nucleotide preference throughout evolution, implying strong conservation. On the other hand, degenerate positions with their flexibility to allow different nucleotides show a higher rate of evolution while maintaining their nucleotide distribution.

Few computational methods exist that consider binding site evolution explicitly. One such approach is proposed by Moses and colleagues as a tool Monkey (Moses *et al* 2004a). Here, given a profile, existing multi-species alignments are scored using a profile-specific model of evolution. Although using an explicit model for TFBS evolution, Monkey relies on existing alignments and hence needs to adopt a heuristic to deal with gaps in the aligned hit locations.

Amongst motif discovery methods, EMnEM (Moses *et al* 2004b), PhyMe (Sinha *et al* 2004), and PhyloGibbs (Siddharthan *et al* 2005) are examples that explicitly consider TFBS evolutionary information. In a different approach, Mustonen *and* Lässig (2005) present and use a model for TFBS evolution in a method that calculates the likelihood of observing a set of aligned sequences under different modes of evolution.

Previously (Bais *et al* 2007), we showed how the simultaneous aligning and annotating strategy helps in highlighting gaplessly aligned TFBS hits by local rearrangements in the alignment. Moreover, we described how the profile related parameters can be derived in a sound statistical framework, thus avoiding ad hoc decisions. In a simulated setting we showed the advantage of using SimAnn (Bais *et al* 2007) over multi-step approaches. Being a simultaneous approach, SimAnn does not require additional strategies to deal with the gaps in the aligned hits. However, there we derive the profile-related parameters by assuming independence between a pair of strings. A log-likelihood ratio test is used to compare between the alternative hypothesis of independent samples of the position-specific letter distribution of the profile versus the null hypothesis of independent samples from the background distribution. Such a scoring disregards the evolutionary relationship between the binding sites and relies more on the individual hit scores. Depending on the quality of profile the score of a pair of strings may be unduly penalized or favored. It is hence desirable to take the background evolution into account, while at the same time model the evolutionary characteristics of the profile.

In this article, we address this issue of considering TFBS evolution for searching conserved TFBSs. We demonstrate how a pair of binding sites can be treated as evolutionarily related and scored accordingly in a statistical framework. While the alignment algorithm remains unchanged, the parameter derivation is now modified. Instead of scoring a

pair of strings based on the independence assumption, we now consider the scenario that the first string is sampled from the profile and then evolved to the second using a profile-based evolutionary model. Predicted hit-pairs are "conserved" binding sites – gaplessly aligned, evolutionarily related profile instances. The parameter calculations are described using two evolutionary models employed previously for modelling TFBS evolution. The additions are implemented in Perl and come together with the new version of SimAnn (eSimAnn) which is available from the authors on request.

In the following, we recapitulate the SimAnn algorithm and scoring details. We then describe in a general setting, how an evolutionary model for TFBSs can be explicitly incorporated into the SimAnn framework. Using two models – the Felsenstein 1981 (F81) (Felsenstein 1981) and the Halpern-Bruno (HB) (Halpern and Bruno 1998), both used previously for modelling the position-specific evolution of TFBSs – we demonstrate how each aspect of the scoring scheme can be estimated in a statistical setting. Next, we provide an initial validation of the proposed approach in a simple simulated setting by comparing eSimAnn with SimAnn. Finally, we provide a real-data evaluation of eSimAnn using experimentally verified binding site data provided by Lenhard *et al* (2003). Here, we compare eSimAnn with the multi-step tool Monkey (Moses *et al* 2004a) which also considers TFBS evolutionary properties.

2. Background – SimAnn algorithm

The underlying idea in SimAnn is to allow for the possibility that a stretch of l nucleotide pairs be annotated either as l consecutive substitutions or as a "pair-profile" hit. This implies that along with the standard alignment substitutions, SimAnn allows for additional pair-profile states. The algorithm is a simple extension of the Smith-Waterman (SW) algorithm (Smith and Waterman 1981).

Briefly, given a profile $P = (P_l \ldots P_l)$ representing binding sites of length l of a factor, and a pair of sequences x and y, we wish to find conserved instances of P while aligning the two sequences. The standard recursion rules of the SW algorithm for filling the dynamic programming matrix M are accordingly modified as:

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

where s and g represent the substitution scoring matrix and the gap cost respectively. The profile scoring array PSA assigns real-valued scores to every pair of strings of length l depicting how well a pair of strings represents the profile. The profile penalty *pen* helps in maintaining a balance between the alignment scenarios: l substitutions or one pair-profile hit.

The alignment is generated by usual traceback, where besides standard substitutions and indels, pair-profile hits are also incorporated. We assume that s is given as the usual log-odds substitution scoring matrix comparing the probability that a pair of nucleotides is related (given by p(u, v)) versus being independently sampled from the background (= $\pi(u)\,\pi(v)$).

Until now, the profile scoring array PSA is calculated for every pair of strings u, v of length l, as a log-likelihood ratio comparing the possibility that (u, v) are independent samples of the profile P to that of each u_i, v_i being independently sampled from the background distribution π. Following the lines of Rahmann *et al* (2003), beginning from a position-specific count matrix we formulate a regularized position-specific probability matrix or profile (P) and a position-specific scoring matrix (PSSM). The final PSA score of a pair of strings can be then calculated as:

$$\begin{aligned} \mathrm{PSA(u,v)} &:= \log\left(\frac{\mathrm{P(u)P(v)}}{\Pi_{i=1}^{l}\pi(u_i)\pi(v_i)}\right) \\ &=: \mathrm{PSSM(u)} + \mathrm{PSSM(v)}. \end{aligned} \tag{2}$$

The profile penalty helps in balancing between the two scenarios: pair-profile hit versus l substitutions. If PSA (u, v) – pen > $\Sigma_{i=1}^{l}$ s (u_i, v_i), then the corresponding stretch in the alignment is assigned to the former rather than the latter, which leads to:

$$\mathrm{LLR}_{\mathrm{P}^2,\mathrm{P}^l}(\mathrm{u,v}) := \log\frac{\mathrm{P(u)P(v)}}{\prod_{i=1}^{l}\mathrm{p}(u_i, v_i)} > pen \tag{3}$$

Hence, comparing the two score distributions leads to a log-likelihood ratio (LLR) test, where the profile penalty can be chosen based on the desired type I and type II error levels. Here, the distribution which arises by sampling two independent strings from the profile is denoted by P^2, and that which arises from independently sampling l evolutionarily related letter pairs from p by p^l. An example choice is the level α type I cutoff where the profile penalty is chosen such that the $\mathrm{P}_{p^l}(LLR_{p^2,\,p^l}(\mathrm{u,v}) > pen)$ is smaller than α. In other words, the type I error probability is less than α. The exact score distributions under the pair-profile model and the background evolutionary model can be calculated using convolution techniques (*see* Rahmann *et al* 2003). We now describe how this approach can be further extended to explicitly consider the TFBS evolutionary properties. As we will see, while the dynamic programming algorithm

remains unchanged, the scoring parameter derivations are modified.

3. Methods

We begin with a description of the general procedure in which any evolutionary model can be incorporated into the SimAnn framework for modelling TFBS evolution. Next, we demonstrate through two evolutionary models, how such an incorporation can be carried out in practice.

3.1 *General procedure*

An evolutionary model for DNA is usually described by a 4-by-4 instantaneous rate matrix Q =($q(u, v)$), $u, v \in$ A,C,G,T (*see* Hillis *et al* 1996, for textbook reference). The off-diagonal entries q (u, v), $u \neq v$ give the instantaneous rate of change from nucleotide u to v, while the diagonal entries are set such that rows sum to 0 each. The transition probability matrix with entries P(t) = ($p_{uv}(t)$) is then calculated as P(t) = exp(Qt), given the rate matrix Q and time t. The entry $p_{uv}(t)$ gives the probability that nucleotide u is replaced by v in time t. We assume that the background distribution π is given by the uniform distribution.

Recall that the substitution scoring matrix scores s (u, v) for background sequence alignment (eq.1) are usually calculated as scaled log-likelihood ratios of observing (u, v) as an evolutionarily related pair versus independent samples from the background distribution (Durbin *et al* 1998). Similarly, equipped with an evolutionary model, PSA(u, v) can now be calculated as the log-likelihood ratio of observing the pair u, v as evolutionarily related binding sites represented by the profile P versus each u_i, v_i sampled independently from the background distribution π. The PSA score at a position i can then be calculated by considering the corresponding position-specific letter distribution of the profile and the evolutionary distribution at that position. Let $\rho = \rho^1 (u_1, v_1, t) \ldots . \rho^l (u_l, v_l, t)$ give the position-specific time-dependent transition probabilities under the profile, then the PSA score at a position i in the profile is given by:

$$\mathrm{PSA}(u_i, v_i) := \log\left(\frac{P^i(u_i)\rho^i(u_i, v_i, t)}{\pi(u_i)\pi(v_i)}\right).$$

Hence the score of u_i, v_i at position i is calculated by comparing the probability of observing u_i in the first string and then evolving it according to ρ^i to v_i.

Comparing the above with eq. (2), we can see that the incorporation of binding site relatedness into the score derivation leads to a loss of the simple additive form in eq. 2. At the same time, the above derivation requires additional rate parameters. For a first approach to estimate these parameters, given a substitution scoring matrix, we adopt the simple strategy of assuming an evolutionary model like the Jukes-Cantor (Jukes and Cantor 1969) for background sequence evolution (details in Appendix).

The profile penalty *pen* now compares between the two alternatives of u, v being evolutionarily related samples of P versus each (u_i, v_i) being a standard substitution. If we denote the former by p′, then at position i, $P'^i(u_i, v_i) = P^i(u_i)\rho^i(u_i, v_i, t)$. This leads to a log-likelihood score formulated using the two distributions p′ and p:

$$\mathrm{LLR}_{p'^l, p^l}(\mathrm{u}, \mathrm{v}) := \log \prod_{i=1}^{l} \frac{p'(u_i, v_i)}{p(u_i, v_i)} > pen\,. \qquad (4)$$

Hence, whereas previously (eq. 3) we compared the distribution under independent sampling of profile versus that under background evolution, here we compare the distribution under TFBS evolution versus that under background evolution. The exact distributions under p′ and p can again be derived using the method of Rahmann *et al* (2003) and the profile penalty chosen according to desired type I and type II error levels.

After providing the generic approach of incorporating any evolutionary model for TFBS evolution, we now describe how two evolutionary models, used commonly for modelling the position-specific evolutionary properties of TFBSs, can be integrated into the approach. It should be stressed here that while both models provide a better approach to modelling position-specific evolution in binding sites as compared to models that treat all posi-tions similarly, both rely on simplifying assumptions. Nevertheless, both provide a more realistic representation of binding site evolution, as shown by Moses *et al* (2004b). We begin with the Felsenstein 1981 model because of its simplicity and ease of incorporation into the SimAnn framework.

3.2 *The Felsenstein 1981 model*

To ensure that each position in a profile is treated differently with regards to evolutionary characteristics, for an initial choice we adapted the Felsenstein 1981 model (F81) (Felsenstein 1981). According to this model, the probability of a substitution is proportional to the stationary distribution of the incoming nucleotide. Setting the stationary distribution at each position i to the position-specific letter distribution under the profile P^i, the model respects the initial base composition through position-specific substitution rates. The position-specific transition probabilities at a position i are then given by:

$$\rho^i(u_i, v_i, t) = e^{-\mu t}\delta_{u_i v_i} + (1 - e^{-\mu t})\,P^i(v_i) \quad \forall\ u_i, v_i\,, \qquad (5)$$

where μ is the rate of mutations per site and δ is the Kronecker delta function with $\delta_{u,v}$=1 if (u = v) and 0 otherwise. Inserting

the above probabilities into eq. 4 and rearranging, the PSA score for the pair of strings is given as:

$$\mathrm{PSA(u,v)} := \mathrm{PSSM(u)} + \mathrm{PSSM(v)} + \sum_{i=1}^{l} \log\left[\left(\frac{\delta_{u_i,v_i}}{\mathrm{P}^i(v_i)} - 1\right) e^{-\mu t} + 1\right].$$

Note how as time goes to infinity, this reduces to the simple additive form in the case of independent scoring in eq. 2. However, when $u_i \neq v_i$; the contribution of the profile letter distribution in the additional term is lost and the score depends purely on the evolutionary rate. While considering a rate slower than background sequence has been proposed (Moses *et al* 2004b) it does not fully reflect the profile conservation properties at each position. Currently, we use the rate as derived from the substitution scoring matrix for background sequence (see Appendix).

3.3 *The Halpern Bruno model*

According to this model, a position-invariant mutation rate is combined with a position-dependent fixation rate to yield position-specific mutation rates $q^i(u, v)$. Briefly, the mutation rate at a position i in the profile is given as the following proportionality:

$$q^i(u_i, v_i) \propto q_b(u_i, v_i) \times \frac{\ln x}{1 - 1/x}, \tag{6}$$

where

$$x = \frac{\mathrm{P}^i(v_i) q_b(v_i, u_i)}{\mathrm{P}^i(u_i) q_b(u_i, v_i)},$$

and $q^b(u, v)$ gives the background evolutionary model. If $x = 1$, then the rate is equal to the background mutation rate $q^b(u, v)$. Hence, the model suitably reflects the position-specific evolution in the profile whereby degenerate positions mutate more while non-degenerate positions are more conserved. Given the rate matrix, the transition probabilities at each position can again be derived by exponentiating the product of time and rate matrix. For a more detailed discussion on the use of the HB model for TFBS evolution (*see* Moses *et al* 2004a,b).

The current implementation of SimAnn supports Perl scripts for modelling TFBS evolution using the F81 model as well as the Halpern-Bruno model, where the rate parameters are calculated using the Jukes-Cantor model for background sequence evolution.

4. Results

4.1 *Simulation analysis*

To demonstrate the feasibility of the proposed extension, we adopted a simulation strategy similar to that used in Bais *et al* (2007). The main difference is now we consider evolutionarily related motif samples instead of independent samples. The embedded motifs are generated by evolving a sampled instance according to either the F81 or the HB model to yield the second motif in a pair. Information about the true motif locations as well as the true alignment are stored for analysis. Two distances and two profiles of medium and poor quality, respectively are considered (*see* Appendix). Both SimAnn and eSimAnn, where the PSA is derived using the respective evolutionary model, are run on each pair using a wide range of profile penalties. At each penalty, the predicted hit pairs are compared with the true locations to retrieve true and false positives (TPs and FPs). These are used to plot receiver operator characteristic (ROC) curves. Through the ROC curves we also assess how theoretically calculated profile penalty in the eSimAnn case performs in terms of true and false positive rates. Details of the simulation setting as well as parameters are presented in Appendix. Focussing first on the use of Halpern-Bruno model, Figures 1 and 2 show the ROC curves for medium and poor quality profiles (Transfac Ids M00690 and M00395) at two distances. Under a good quality profile, SimAnn (blue) and eSimAnn (black) perform very similar (data not shown). However, as the profile quality deteriorates, eSimAnn considerably outperforms the former. This is also the case with increase in distance, where although both methods suffer eSimAnn has a clear advantage. In each case, the ROC curves of eSimAnn rise steeper and higher than those of SimAnn. Similar results hold under the F81 model (figures 3 and 4). A point to stress here is that the fall in the curves of both SimAnn and eSimAnn at extremely low profile penalties arises due to the fact that the algorithm does not predict overlapping hits. Hence, when the profile penalty decreases extremely, it tries to fill the alignment with as many non-overlapping instances of pair profiles as possible, and thus loses the predictions made correctly at higher penalties. The profile penalties highlighted on the eSimAnn curves illustrate the validity of the theoretical calculations.

4.2 *Real data – comparison with a multi-step approach*

Majority of TFBS prediction methods that use conservation information are multi-step, thus relying on the correctness of an existing alignment. Only few additionally consider binding site evolution characteristics. Of interest to us is the tool Monkey (Moses *et al* 2004a) which although multi-step, is most similar to our approach. Before going further, certain features of Monkey need to be mentioned here.

Based on putative TFBS hits on a reference sequence, Monkey scans corresponding regions of other sequences in a multiple alignment to identify putative conserved TFBS hits. For each putative conserved hit, Monkey outputs a *p*-value estimate based on the Halpern-Bruno model. As with

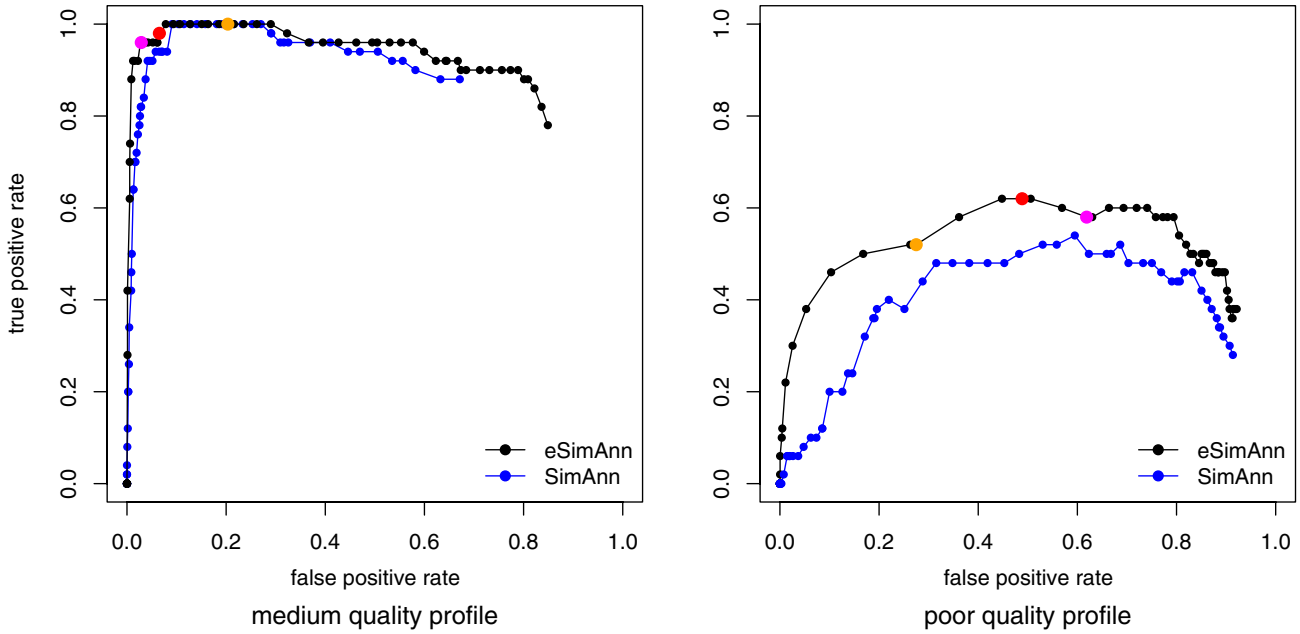

Figure 1. Halpern-Bruno model – Comparison of eSimAnn (black) with SimAnn (blue) at a close evolutionary distance of 0.1 and 2 and using profiles of medium (left) and poor (right) quality, respectively. The ROC curves show the true and false positive rates with varying profile penalty cutoffs. On the eSimAnn ROC curves the theoretically derived profile penalties are highlighted. (Orange, type I error penalty at level 0.05; red, balanced penalty; magenta, type II error penalty at level 0.05.)

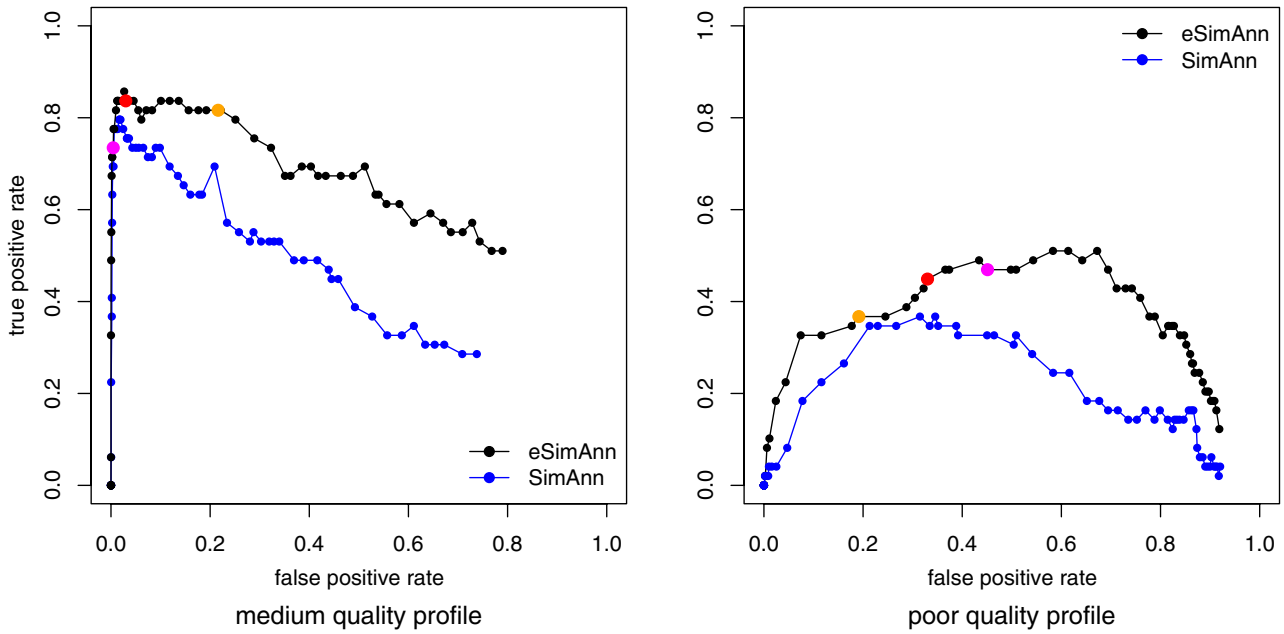

Figure 2. Halpern-Bruno model – Comparison of eSimAnn (black) with SimAnn (blue) at a close evolutionary distance of 0.5 and 2 and using profiles of medium (left) and poor (right) quality, respectively. On the eSimAnn ROC curves the theoretically derived profile penalties are highlighted. (Orange, type I error penalty at level 0.05; red, balanced penalty; magenta, type II error penalty at level 0.05.)

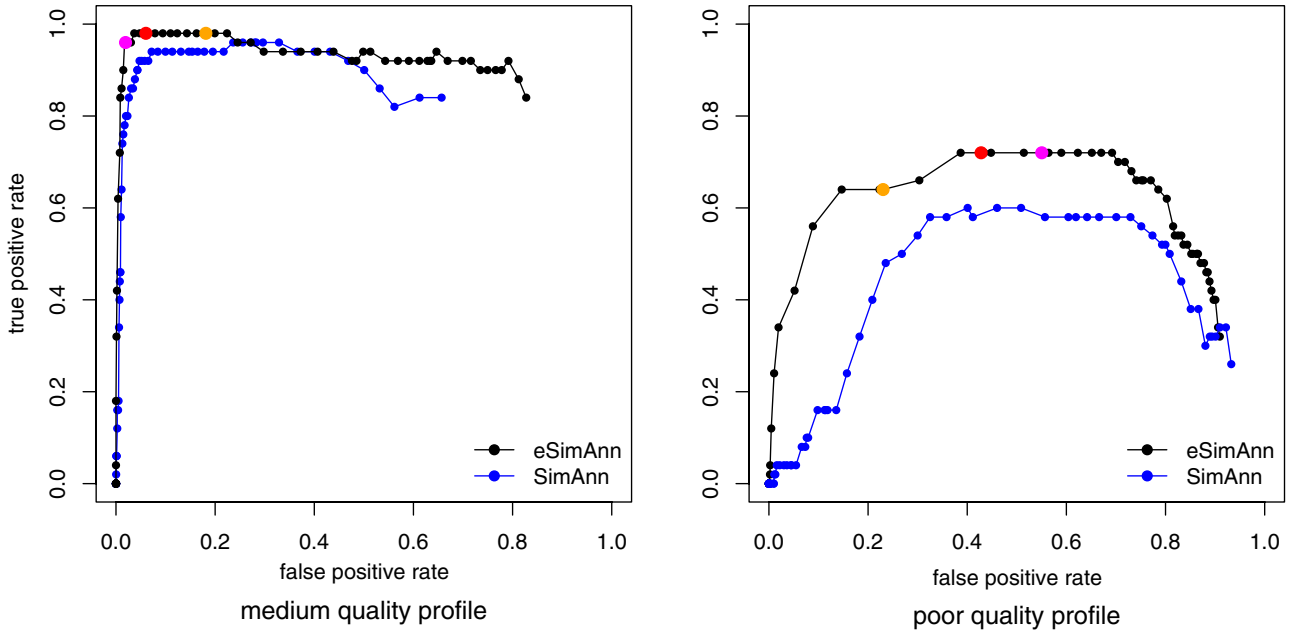

Figure 3. Felsenstein 1981 model – Comparison of eSimAnn (black) with SimAnn (blue) at a close evolutionary distance of 10 and 2 and using profiles of medium (left) and poor (right) quality, respectively. The ROC curves show the true and false positive rates with varying profile penalty cutoffs. On the eSimAnn ROC curves the theoretically derived profile penalties are highlighted. (Orange, type I error penalty at level 0.05; red, balanced penalty; magenta, type II error penalty at level 0.05.)

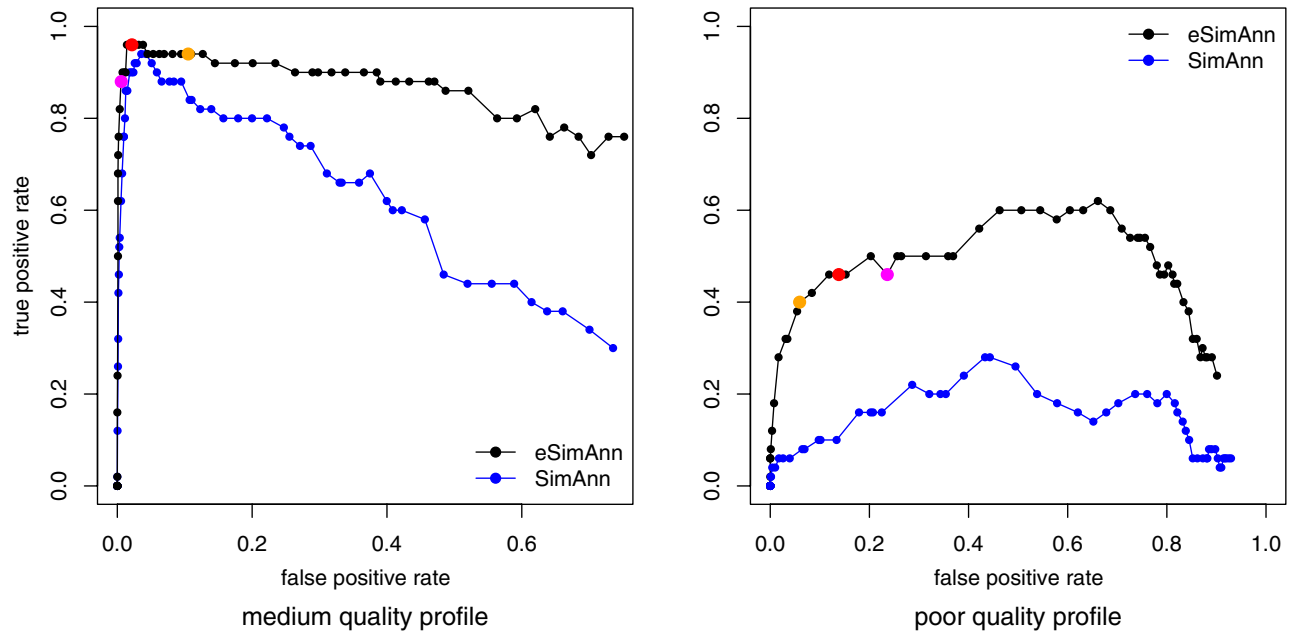

Figure 4. Felsenstein 1981 model – Comparison of eSimAnn (black) with SimAnn (blue) at a close evolutionary distance of 50 and 2 and using profiles of medium (left) and poor (right) quality, respectively. On the eSimAnn ROC curves the theoretically derived profile penalties are highlighted. (Orange, type I error penalty at level 0.05; red, balanced penalty; magenta, type II error penalty at level 0.05.)

other multi-step methods, Monkey needs to deal with gaps in the aligned hit locations. To this end, it uses a heuristic that conservatively allows some gaps while disallowing too many.

We used a testset of experimentally verified binding sites in human-mouse sequences provided by Lenhard *et al* (2003). Removing those examples that contain ambiguous

characters (Ns) results in a set of 98 experimentally verified binding sites in human-mouse orthologous sequences. Count matrices provided here were converted to the respective regularized position-specific probability and scoring matrices using the methods of Rahmann *et al* (2003).

To run eSimAnn, we used the default HOXD70 substitution scoring matrix, shown to be suitable for human-mouse comparisons (Chiaromonte *et al* 2002). For comparisons with Monkey which uses the HB model, the same is used to calculate the profile-related parameters. For background evolution, the Jukes-Cantor model with uniform background frequencies is used. We ran Monkey on alignments generated using ClustalW with the default parameters. The same distance as estimated from the HOXD70 matrix and Jukes Cantor model is used. The background frequencies were set to uniform. Since, in contrast to Monkey, SimAnn does not predict overlapping hits, we calculated the true and false positive predictions at nucleotide level. A prediction is called a true positive if it overlaps by more than half the length of the shorter of either the count matrix or the known site [similar to Lenhard *et al* (2003)]. Multiple overlapping true hits are ignored and the number of correctly predicted bases is limited to the length of the respective matrix.

Monkey outputs a list of putative hits with the associated pvalues making it non-trivial to decide on the appropriate pvalue threshold. To deal with this, we consider a range of pvalue thresholds. We ran eSimAnn with profile penalties calculated for each of these p value thresholds (p value levels) and plotted the resulting proportion of true and false predictions under both methods (Figure 5). As can be seen, eSimAnn performs comparably to Monkey,

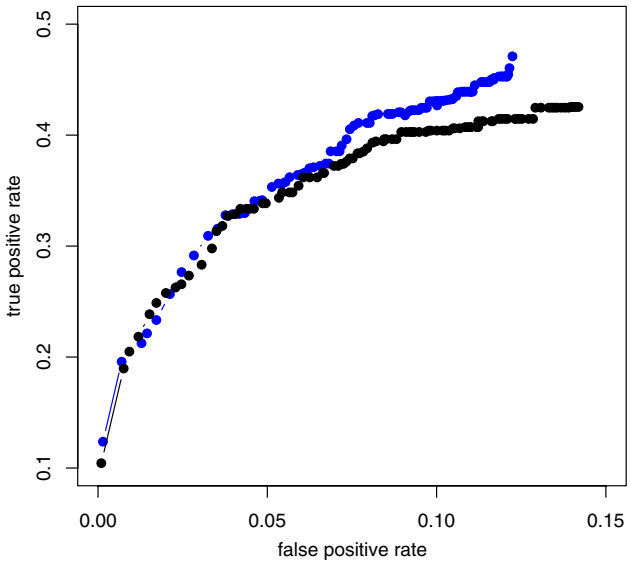

Figure 5. True versus false positive rates on real testset of human-mouse sequences with experimentally verified binding sites for eSimAnn (black) and Monkey (blue).

showing slightly better performance at lower p value thresholds.

5. Discussion

Previously (Bais *et al* 2007), we had introduced a simultaneous alignment and annotation approach which integrates profiles into the alignment generation step yielding annotated alignments. We proposed and validated a statistically motivated choice for the parameters eliminating the need of ad hoc cutoff choices. The simultaneous approach allows for local shuffling of gaps to highlight gaplessly aligned TFBS hits. However SimAnn scores individual hits in a pair independently, thereby ignoring the TFBS evolutionary properties. In this article, we demonstrated how TFBS evolution can be incorporated into the framework of annotated alignments.

We first provided a general outline of how an evolutionary model suitable for reflecting binding site evolution, can be used to derive profile parameters in eSimAnn. To describe the procedure in practice, we used two models employed previously for TFBS evolution. Using a controlled setting we studied the difference in performance of eSimAnn and SimAnn, simultaneously verifiying the theoretical derivations for the parameter choice. Finally, using a real dataset and the existing multi-step tool Monkey, we showed that eSimAnn performs comparably well. It should be stressed that Monkey can use multiple alignments while eSimAnn is a pairwise alignment algorithm. However, being a multi-step approach, Monkey needs to resort to heuristics for dealing with gaps in the aligned hit locations. In contrast, SimAnn (eSimAnn) is an optimal alignment algorithm where the gaps are locally re-arranged on the fly to bring forth conserved TFBSs. This also means that it is not suitable for large scale analysis of arbitrary sequences. The applicability of eSimAnn as purely an alignment tool can be an interesting direction for further research. Here the annotated alignments generated by eSimAnn could be used with a multi-step tool like Monkey. In such a scenario, the multi-step tool would face less hinderance due to gaps in the aligned TFBS hits.

Acknowledgements

AB wishes to acknowledge Dennis Kostka for helpful discussions and critical comments on the manuscript.

Appendix

Estimating the rate parameters

Given an appropriate substitution scoring matrix s, we retrace the rate parameters by assuming that the background

sequences evolved according to a simple evolutionary model. We use the Jukes-Cantor (JC) model (Jukes and Cantor 1969) for simplicity, although more sophisticated models can be similarly employed. Given a substitution scoring matrix s, we can write the probability that a pair of nucleotides is related in terms of the log-likelihood scores:

$$P(u, v) = (e^{s(u,v)}) * \pi(u)\,\pi(v).$$

Using the transition probabilities as derived from the JC model, we also get:

$$\mathrm{P}(u,v) = \pi(u)\left[e^{-\mu t}\delta_{uv} + (1-e^{-\mu t})\,\pi(v)\right] \quad \forall\ u,v\,.$$

Hence, the unknown parameter pair μt, where μ is the mean instantaneous substitution rate and t is the time elapsed, can be estimated from the above two equations.

Simulated setting

We used the software program CisEvolver (Pollard *et al* 2006) for simulating sequence pairs with TFBS evolution modelled according to the HB model. For simulations with the F81 model, we used the software program Rose version 1.3 (Stoye *et al* 1998). In both cases, we generate 50 pairs of evolutionarily related sequences with average lengths of 500 and Jukes-Cantor as the background evolutionary model. Two distance settings of 0.1, 0.5 for CisEvolver and 10, 50 for Rose are used. We use balanced quality as described in Rahmann *et al* (2003) as a measure of profile quality. Similar to Bais *et al*. (2007), count matrices are retrieved from Transfac (Matys *et al* 2003) and are M00395 (poor quality, 0.199) and M00690 (medium quality, 0.622). Gaps are not allowed in the motif locations. In both approaches, the sequence pairs were taken to be at the leaves of a simple depth one binary tree, with branch lengths proportional to the distance. Background frequencies were set to uniform.

In CisEvolver, an ancestor sequence with a single binding site implanted at a random position is evolved to a desired branch length, yielding a pair of evolutionarily related sequences. As mentioned earlier, the implanted motif evolves according to the HB model and we store the true locations of the evolved motifs. For the file for indel frequencies ("indel lengths file" in CisEvolver), we use the example file provided in the CisEvolver package, but set the relative indel rate to 0.05 for a reasonable proportion of gaps in the true alignments.

In Rose, for background sequence evolution we use the default DNA parameters with the indel thresholds at 0.002. Motifs are evolved according to the F81 model – at each position the stationary distribution is set to the corresponding position-specific letter distribution of the profile. A random position in a sequence is chosen and a motif sampled from the respective profile is implanted. At the equivalent position in the true alignment the evolved motif is implanted in the second sequence.

To run SimAnn and eSimAnn, we derived the respective substitution scoring matrix for each distance under the Jukes-Cantor model and uniform background frequencies. The gap costs were estimated as before (Bais *et al* 2007). Briefly, for both CisEvolver and Rose, we generate sequence pairs at a fixed evolutionary distance. Next, we realign the sequence pairs, using a wide range of gap cost settings, with gap extension cost set to 1/10 of gap opening cost. The proportion of gaps in the true and recomputed alignments is compared to determine the optimal gap open cost. For the pair-profile parameters, SimAnn is run with PSA calculated using the independence assumption, while eSimAnn is run with that calculated using the corresponding evolutionary model. Finally, a prediction is called a true positive only if it overlaps exactly with the true site locations on both the sequences.

References

Bais A S, Grossmann S and Vingron M 2007 Simultaneous alignment and annotation of cis-regulatory regions; Bioinformatics 23 e44–49

Berg J, Willmann S and Lässig M 2004 Adaptive evolution of transcription factor binding sites; *BMC Evol. Biol* **4** 42

Chiaromonte F, Yap V B and Miller W 2002 Scoring pairwise genomic sequence alignments; *Pac. Symp. Biocomput*. 115–126

Dermitzakis E T and Clark A G 2002 Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover; *Mol. Biol. Evol*. **19** 1114–1121

Durbin R, Eddy S, Krogh A and Mitchison G 1998 *Biological sequence analysis* (Cambridge: Cambridge University Press)

Felsenstein J 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach; *J. Mol. Evol* . **17** 368–376

Gerland U and Hwa T 2002 On the selection and evolution of regulatory DNA motifs; *J. Mol. Evol*. **55** 386–400

Halpern A L and Bruno W J 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies; *Mol. Biol. Evol*. **15** 910–917

Hillis D M, Moritz C and Mable B K 1996 *Molecular systematics* (Sunderland, MA: Sinauer Associates)

Jukes T H and Cantor C R 1969 Evolution of Protein Molecules; in *Mamalian protein molecules* (ed.) H N Munro (New York: Academic Press) vol. 3, pp 21–132

Kotelnikova E A, Makeev V J and Gelfand M S 2005 Evolution of transcription factor DNA binding sites; *Gene* **347** 255–263

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N and Wasserman W W 2003 Identification of conserved regulatory elements by comparative genome analysis; *J. Biol*. **2** 13

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, Palsson A, Alekseeva E, Bergman C M, Nathan J and Kreitman M 2005 Functional evolution of a cis-regulatory module; *PLoS Biol*. **3** e93

MacIsaac K D and Fraenkel E 2006 Practical strategies for discovering regulatory DNA sequence motifs; *PLoS Comput. Biol*. **2** e36

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D *et al* 2003 TRANSFAC: transcriptional regulation, from patterns to profiles; *Nucleic Acids Res.* **31** 374–378

McCue L A, Thompson W, Carmack C S and Lawrence C E 2002 Factors influencing the identification of transcription factor binding sites by cross-species comparison; *Genome Res*. **12** 1523–1532

Moses A, Chiang D, Pollard D, Iyer V and Eisen M 2004a Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model; *Genome Biol*. **5** R98

Moses A M, Chiang, D Y, Kellis M, Lander E S and Eisen M B 2003 Position specific variation in the rate of evolution in transcription factor binding sites; *BMC Evol. Biol*. **3** 19

Moses A M, Chiang DY and Eisen M B 2004b Phylogenetic motif detection by expectation maximization on evolutionary mixtures; *Pac Symp. Biocomput*. 324–335

Mustonen V and Lässig 2005 Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies; *Proc. Natl. Acad. Sci. USA* **102** 15936–15941

Pollard D A, Moses A M, Iyer V N and Eisen M B 2006 Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments; *BMC Bioinformat*. **7** 376

Rahmann S, Müller T and Vingron M 2003 On the power of profiles for transcription factor binding site detection; *Stat. Appl. Genet. Mol. Biol*. **2** Article 7

Siddharthan R, Siggia E D and van Nimwegen E 2005 PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny; *PLoS Comput. Biol*. **1** e67

Siggia E D 2005 Computational methods for transcriptional regulation; *Curr. Opin. Genet. Dev*. **15** 214–221

Sinha S, Blanchette M and Tompa M 2004 PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences; *BMC Bioinformat*. **5** 170

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

Stormo G D 2000 DNA binding sites: representation and discovery; *Bioinformatics* **16** 16–23

Stoye J, Evers D and Meyer F 1998 Rose: generating sequence families; *Bioinformatics* **14** 157–163

Sui S J H, Mortimer J R, Arenillas D J, Brumm J, Walsh C J, Kennedy B P and Wasserman W W 2005 oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes; *Nucleic Acids Res*. **33** 3154–3164

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

Wittkopp P J 2006 Evolution of cis-regulatory sequence and function in Diptera; *Heredity* **79** 139–147

*e*Publication: 5 July 2007