BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 3 2008, pages 350–357
doi:10.1093/bioinformatics/btm610

Sequence analysis

Natural similarity measures between position frequency matrices with an application to clustering

Utz J. Pape[1,2,*], Sven Rahmann[3,4] and Martin Vingron[1]

[1]Computational Biology, Max Planck Institute f. Molecular Genetics, Ihnestr. 73, 14195 Berlin, [2]Mathematics and Computer Science, Free University of Berlin, Takustr. 9, 14195 Berlin, [3]COMET group, Genome Informatics, Universität Bielefeld, 33594 Bielefeld and [4]Bioinformatics for High-Throughput Technologies, Computer Science 11, Dortmund University, 44221 Dortmund, Germany

Received on August 30, 2007; revised on December 5, 2007; accepted on December 6, 2007

Advance Access publication January 2, 2008

Associate Editor: Alfonso Valenica

ABSTRACT

Motivation: Transcription factors (TFs) play a key role in gene regulation by binding to target sequences. *In silico* prediction of potential binding of a TF to a binding site is a well-studied problem in computational biology. The binding sites for one TF are represented by a position frequency matrix (PFM). The discovery of new PFMs requires the comparison to known PFMs to avoid redundancies. In general, two PFMs are similar if they occur at overlapping positions under a null model. Still, most existing methods compute similarity according to probabilistic distances of the PFMs. Here we propose a natural similarity measure based on the asymptotic covariance between the number of PFM hits incorporating both strands. Furthermore, we introduce a second measure based on the same idea to cluster a set of the Jaspar PFMs.

Results: We show that the asymptotic covariance can be efficiently computed by a two dimensional convolution of the score distributions. The asymptotic covariance approach shows strong correlation with simulated data. It outperforms three alternative methods. The Jaspar clustering yields distinct groups of TFs of the same class. Furthermore, a representative PFM is given for each class. In contrast to most other clustering methods, PFMs with low similarity automatically remain singletons.

Availability: A website to compute the similarity and to perform clustering, the source code and Supplementary Material are available at http://mosta.molgen.mpg.de

Contact: utz.pape@molgen.mpg.de

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

1 INTRODUCTION

Detection of binding sites is a crucial task in deciphering gene regulatory networks (Wasserman and Sandelin, 2004). Binding sites of transcription factors (TFs) are often represented as Position Frequency Matrices (PFMs) as introduced by Stormo *et al.* (1982). Based on this representation, first studies of the affinity of TFs to DNA sequences were performed (Schneider *et al.*, 1986; Staden, 1984; Stormo *et al.*, 1982). Stormo (2000) shows that the score of a sequence is proportional to its binding energy. Score calculation has been refined to improve discrimination between sites and non-sites (Berg and von Hippel, 1987; Hertz *et al.*, 1990; Stormo and Hartzell, 1989). The threshold for the score can be calculated such that the probability for a false positive (type I error) is controlled at a certain level (Rahmann *et al.*, 2003).

Many computational tools deal with *ab initio* discovery of new PFMs on a set of related sequences (Tompa *et al.*, 2005). Since there is no best method, several programs are usually applied resulting in a redundant set of PFMs. Furthermore, the methods might discover PFMs similar to known PFMs. Therefore, either similar PFMs should be removed or merged into a new PFM. Thus, an appropriate similarity measure for PFMs is required.

Most similarity measures consider PFMs as probability distributions. Hence, the distance between the distributions is used as dissimilarity measure. Due to the position independence of PFMs, the comparison is done column-by-column which has been shown to work well (Liu *et al.*, 1990). The Pearson correlation coefficient which has been shown to be more effective than other methods (Pietrokovski, 1996) is widely used. Wang and Stormo (2003) describe the average log-likelihood ratio method. Schones *et al.* (2005); Kielbasa *et al.* (2005) calculate the independence of the columns of two PFMs using the χ^2 statistic (Fleiss *et al.*, 2003). The Kullback-Leibler distance is also often used (Aerts *et al.*, 2003; Roepcke *et al.*, 2005). The Tomtom algorithm (Gupta *et al.*, 2007) can use any of these measures to compute a null distribution of similarity scores to obtain *P*-values. An additional measure described by Kielbasa *et al.* (2005) does not compute the distance between the PFM distributions but the correlation between the scores of the PFMs on a given sequence.

Suzuki and Yagi (1994) show that TFs of the same family share similarity in the binding profile. The Familial Binding Profile (FBP) is a generalized binding profile capturing this core motif of a family of TFs (Sandelin and Wasserman, 2004). Several approaches perform clustering of TFs into families based on a Bayesian learning algorithm (Narlikar and Hartemink, 2006) and unsupervised neural network (Mahony *et al.*, 2005). Others use as a metrics ungapped local motif

*To whom correspondence should be addressed.

alignments (Sandelin and Wasserman, 2004) and similarity measures (Kielbasa *et al*., 2005). A comparison of DNA sequence based approaches is presented by Mahony *et al*. (2007).

In spite of the wealth of literature on this topic, to date there is no 'natural' definition of the similarity of two PFMs. Here we propose what we think is a natural similarity measure: Two PFMs should be regarded as similar when they describe similar binding sites. In this case, they yield a high number of overlapping hits on a random sequence. Hence, the number of hits on the sequence is correlated between both PFMs. Considering the number of hits for a PFM on a random sequence as a random variable, the correlation is captured in the covariance between the random variables of two PFMs. We normalize the covariance by the sequence length and compute the asymptotic covariance for the sequence length approaching infinity. Furthermore, we introduce a related measure based on log-odd scores for the maximum overlap probability for the clustering.

The covariance approach is related to the score correlation method (Kielbasa *et al*., 2005): The covariance capturing the tendency of overlapping hits is derived using the 2D joint score distributions for each possible overlap between both PFMs. The two dimensions of the joint distributions correspond to the score distributions of the two PFMs. On the one hand, the probability of an overlapping hit is the quantile of the joint score distribution with both scores greater or equal than the corresponding thresholds. On the other hand, the score correlation method approximates the correlation between both score distributions. A higher correlation of the scores is related to a higher joint probability of scores greater or equal than the thresholds. Therefore, both approaches are based on similar ideas. However, the new approach presented here does not use an approximation for the score correlation and it naturally summarizes the possible overlap positions by computing the covariance.

As mentioned above, Gupta *et al*. (2007) developed the Tomtom algorithm to compute the null distribution of similarity scores. Although we use a similar algorithm, we do not compute the null distribution of similarity scores but of motif scores based on the PFMs. Hence, we circumvent the arbitrary choice of a column-by-column similarity measure and, instead, we can use the covariance for summarization instead of a minimum *P*-value statistic.

We show the performance of the new approach by a simulation. We use the ratio of overlapping and non-overlapping hits in simulated sequences to obtain a similarity between pairs of PFMs. A generated PFM family is used for comparison with the χ^2 test which performed best in Schones *et al*. (2005), the Kullback-Leibler distance (Aerts *et al*., 2003; Roepcke *et al*., 2005), and the best Tomtom approach (Gupta *et al*., 2007) using the Euclidean distance. Since the exact Fisher-Irwin test (Bailey, 1977) is as good as the χ^2 test, we focussed on the latter one. Furthermore, we omit the score correlation (Kielbasa *et al*., 2005) because its performance is similar to the χ^2 test (Kielbasa *et al*., 2005). We also use a subset of Transfac (Matys *et al*., 2003) PFMs and correlate the simulated similarities with our approaches. Finally, we introduce a related similarity measure based on the maximum overlap probability. Since this measure automatically returns the position with the highest overlap probability, we obtain a gapless alignment of the two PFMs. Hence, we can merge the PFMs. Therefore, we apply this measure to cluster a set of class labelled Jaspar (Sandelin *et al*., 2004) PFMs and compare the class of the members for each cluster. We also automatically obtain familial binding profiles for each cluster. The results are compared with the ones from Mahony *et al*. (2007).

In the remainder of the article, we develop the statistics for the computation of the new approaches and shortly review the alternative methods. We also describe the generation of the PFM family for the simulation. Finally, we present a comparison of the approaches, the performance on Transfac PFMs, the clustering of Jaspar PFMs and discuss the impact of the results.

2 METHODS

2.1 Binding sites and words

The PFM is a representation of a TF binding site. This matrix contains the probabilities for each nucleotide at every position. The position specific scoring matrix (PSSM) is given by the log-likelihood ratios of the nucleotide distribution of the PFM and the background probabilities. As background model, we use a symmetric i.i.d. model incorporating the average GC content of the upstream sequence. We restrict ourselves to the GC content instead of base pair composition to make the computation invariant with respect to the choice of the leading strand.

Using the PSSM of a TF A of length n_A, we can assign a score to every position of the sequence with potential binding sites depending on the observed nucleotide. A position j is a hit if the word $a = a_0 a_1 \dots a_{n_A-1}$ at the jth position of the sequence yields a summed score $s_A(a)$ greater or equal to a threshold t_A. We denote this event with an indicator random variable $Y_j^A = 1$. Independent of a given sequence, we can determine for each word a of length n_A its score $s_A(a)$. Each word corresponding to a hit ($s_A(a) \geq t_A$) is called a compatible word of A. The set of all compatible words is denoted by $\mathcal{A}$. We introduce random indicator variables Y_j^a which are 1 if the word at position j of the sequence is a and otherwise 0. Since a hit of TF A at position j occurs if the word at position j of the sequence is in $\mathcal{A}$, we obtain

$$Y_j^A \equiv \sum_{a \in \mathcal{A}} Y_j^a,$$

since hits are necessarily disjoint at each position.

For simplicity, we start with two TFs A and B each with only one compatible word: $|\mathcal{A}| = 1$ and $|\mathcal{B}| = 1$. Furthermore, we ignore the complementary strand for the beginning. Afterwards, we discard this restriction and allow any size of $\mathcal{A}$ and $\mathcal{B}$.

2.2 Asymptotic covariance for words

In this section, we derive the formulas for the asymptotic covariance between the counts of two words $a \in \mathcal{A}$ and $b \in \mathcal{B}$ (Reinert *et al*., 2000; Waterman, 2000). The section follows the presentation in Waterman (2000) with simplified background model H_0 which is defined by an i.i.d. sequence with background frequencies $\pi(\sigma)$ for each letter $\sigma \in \Sigma$. We can compute the probability for an occurrence of a under the background model H_0 by

$$\alpha_a := \mathbb{P}_{H_0}(Y_j^a = 1) = \prod_{\tau=0}^{n_a-1} \pi(a_\tau).$$

The number of counts in a sequence region of length m is $N_a(m) = \sum_{j=0}^{m-n_a} Y_j^a$. Before we can state the asymptotic covariance, we introduce some further notation regarding the overlap between two words a and b.

We define the probabilities $\gamma_{a,b}(k)$ for an overlap of word a with a word b at position k of a. For that, we use the overlap bit $\varepsilon_{a,b}(k)$ which is 1 if the words allow the overlap ($a_k = b_0,\ a_{k+1} = b_2, \ldots, a_{n_a-1} = b_{n_a-k-1}$) and otherwise 0. Without loss of generality, we assume $n_b \geq n_a$ and obtain the overlap probability under the background model H_0:

$$\gamma_{a,b}(k) := \mathbb{P}_{H_0}(Y_j^a = 1,\ Y_{j+k}^b = 1) = \epsilon_{a,b}(k) \cdot \alpha_a \cdot \prod_{\tau=n_a-k}^{n_b-1} \pi(b_\tau).$$

We capture the overlap probabilities for each k in the overlap sum $G_{a,b}$ defined as

$$G_{a,b} := \sum_{k=0}^{n_a-1} \gamma_{a,b}(k).$$

The covariance between the counts of A and B on a sequence of length m is the sum over all covariances between the hit indicator random variables:

$$\operatorname{cov}(N_a(m), N_b(m)) = \sum_{i=0}^{m-n_a} \sum_{j=0}^{m-n_b} \operatorname{cov}(Y_i^a, Y_j^b).$$

Since the covariance between non-overlapping hits is zero, we only have to consider overlapping hits. The covariance for overlapping hits is given by

$$\begin{aligned}\operatorname{cov}(Y_i^a, Y_{i+k}^b) &= \mathbb{E}(Y_i^a \cdot Y_{i+k}^b) - \mathbb{E}(Y_i^a) \cdot \mathbb{E}(Y_{i+k}^b) \\ &= \gamma_{a,b}(k) - \alpha_a \alpha_b,\end{aligned}$$

for $0 \leq k \leq n_a$. Hence, we can express the asymptotic covariance between a and b:

$$\begin{aligned}&\lim_{m\to\infty} m^{-1}\operatorname{cov}(N_a(m), N_b(m)) \\ &\quad = G_{a,b} + G_{b,a} - (n_a + n_b)\alpha_a\alpha_b - \alpha_a\big(\gamma_{a,b}(0) - \alpha_b\big).\end{aligned}$$

The two first terms correspond to the overlap probability of a followed by b and b followed by a. The third term contains the product of the expected values of the two random variables. Lastly, we add the covariance for a and b occurring at the same position.

2.3 Asymptotic covariance for binding sites

As we have already seen before, each TF encodes a set of compatible words. Therefore, we have to generalize the asymptotic covariance to deal with sets of words. Again, we consider two TFs A and B with sets of compatible words $\mathcal{A}$ and $\mathcal{B}$. Obviously, the length of each word is the same within each corresponding set. The probability α_A for a hit of TF A is given by

$$\alpha_A := \mathbb{P}_{H_0}(Y_j^A = 1) = \mathbb{P}_{H_0}\left(\sum_{a\in\mathcal{A}} Y_j^a = 1\right) = \sum_{a\in\mathcal{A}} \alpha_a.$$

The definition of the overlap probabilities for TFs follows the same reasoning. An overlap occurs between TF A and B if any of the words in $\mathcal{A}$ overlap with any of the words in $\mathcal{B}$. Since the events of the indicator random variables Y_j^a for $a \in \mathcal{A}$ and respectively for $\mathcal{B}$ are disjoint for fixed position j, we obtain

$$\gamma_{A,B}(k) := \mathbb{P}_{H_0}(Y_j^A = 1,\ Y_{j+k}^B = 1) = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \gamma_{a,b}(k).$$

The sums iterate over the set of compatible words. In the Algorithm section, we show how to avoid the summations and to compute the overlap probabilities efficiently.

The sum of the overlap probabilities is given by

$$G_{A,B} := \sum_{k=0}^{n_a-1} \gamma_{A,B}(k).$$

Eventually, we have to define the number of hits for a TF. Again, it is the sum of the number of hits for the all compatible words: $N_A(m) := \sum_{a\in\mathcal{A}} N_a(m)$. Hence, we can split up the asymptotic covariance for the TFs into sums of asymptotic covariances of words and then rewrite it with the introduced notation:

$$\begin{aligned}&\lim_{m\to\infty} m^{-1}\operatorname{cov}(N_A(m), N_B(m)) \\ &\quad = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \lim_{m\to\infty} m^{-1}\operatorname{cov}(N_a(m), N_b(m)) \\ &\quad = G_{A,B} + G_{B,A} - (n_A + n_B)\alpha_A\alpha_B - \alpha_A\big(\gamma_{A,B}(0) - \alpha_B\big).\end{aligned}$$

Since we also want to consider reverse complementary overlaps, we have to further extend the asymptotic covariance. Denoting the reverse complementary set of words $\mathcal{A}$ by $\mathcal{A}'$ and the corresponding TF variable by A', the symmetry of the restrictive background model H_0 yields $\alpha_A = \alpha_{A'}$ and correspondingly for B, and $G_{A',B} = G_{A,B'}$, $G_{A',B'} = G_{A,B}$, $\gamma_{A',B}(0) = \gamma_{A,B'}(0)$, $\gamma_{A',B'}(0) = \gamma_{A,B}(0)$. Hence, we obtain the following definition for the similarity between two TFs A and B:

$$\begin{aligned}S(A,B) &:= \lim_{m\to\infty} m^{-1}\operatorname{cov}(N_A(m) + N_{A'}(m), N_B(m) + N_{B'}(m)) \\ &= 2 \cdot \big[G_{A,B} + G_{A',B} + G_{B,A} + G_{B',A}\big] \\ &\quad - 4 \cdot (n_A + n_B)\alpha_A\alpha_B \\ &\quad - 2\alpha_A\big(\gamma_{A,B}(0) + \gamma_{A',B}(0) - 2\alpha_B\big).\end{aligned}$$

2.4 Algorithm

As mentioned above, overlap probabilities $\gamma_{A,B}(\cdot)$ for the compatible sets of words still have to be computed. Unfortunately, the enumeration of compatible words for a TF is NP-hard (Zhang *et al.*, 2007). Therefore, we avoid enumeration but compute $\gamma_{A,B}(\cdot)$ using the 2D score distribution of the PSSMs of TF A and B.

In fact, we have to compute the probability of the joint event of two scores greater than or equal to the threshold since $\gamma_{A,B}(k) = \mathbb{P}_{H_0}(Y_j^A = 1,\ Y_{j+k}^B = 1)$. This means that the score for TF A at position j has to be $s_A \geq t_A$ and correspondingly at position $j+k : s_B \geq t_B$. The two scores induce a 2D distribution. Obviously, both scores are not independent for $0 \leq k \leq n_A$. Since scores are the sum of the position specific scores of the PSSM we can decompose the scores into each pair of positions $j+i$ and $j+k+i$. Then, pairs of scores are independent. Hence, we can use a dynamic programming algorithm for each possible shift.

The dynamic programming approach is often used for the computation of 1D score distributions (Beckstette *et al.*, 2006; Claverie and Audic, 1996; Rahmann, 2003; Rahmann *et al.*, 2003; Staden, 1989; Wu *et al.*, 2000). Extension to two dimensions is reported in Pape *et al.* (2007):

Here, we briefly review the algorithm: We denote the PSSM for TF A by $\Psi_{\kappa,\sigma}^A$. We enlarge $\Psi_{\kappa,\sigma}^A$ with zero to the left and to the right: $\Psi_{\kappa,\cdot}^A := 0$ for $\kappa < 0$ or $\kappa > n_A$ and correspondingly for Ψ^B. Without loss of generality, we assume that a prefix of B overlaps with a suffix of A. The idea is to compute the score distribution of a prefix of length $i+1$ of the TF based on the score distribution of the prefix of length i. Let $Q_i^{(k)}(s_A, s_B)$ denote the probability for a score s_A at the first $i+1$ positions of the TF A and a score s_B at the first $i-k+1$ positions of the TF B. This means TF B is shifted by k position to the right of A. We obtain this probability from the last step for a prefix of length i for a score minus the score for the new nucleotide. Hence, we have to look up the score for each nucleotide and sum the corresponding probabilities.

In the initial step, no nucleotide has been observed. Therefore, we set the probability for a score of 0–1 and for all other scores to zero. We obtain for $0 \le k \le n_A$ and $0 \le i \le n_A + k$

$$\begin{aligned} Q_{-1}^{(k)}(s_A, s_B) &:= \begin{cases} 0 & \text{if} \quad s_A \neq 0 \quad \text{or} \quad s_B \neq 0, \\ 1 & \text{otherwise,} \end{cases} \\ Q_i^{(k)}(s_A, s_B) &:= \sum_{\sigma \in \Sigma} Q_{i-1}^{(k)}(s_A - \Psi_{i,\sigma}^A, s_B - \Psi_{i-k,\sigma}^B) \cdot \pi_\sigma. \end{aligned}$$

After the last step, $Q_{n_A+k}^{(k)}(s_A, s_B)$ contains the probability to observe score s_A starting at position j and score s_B starting at position $j+k$. Since we require scores to be greater or equal to the threshold, we obtain the overlap probability:

$$\gamma_{A,B}(k) = \sum_{s_A \ge t_A} \sum_{s_B \ge t_B} Q_{n_A+k}^{(k)}(s_A, s_B).$$

The overlap probabilities for reverse complementary sequences are obtained similarly by using the correspondingly transformed PSSMs. We also use some speed improvements similar to those of Beckstette *et al.* (2006) which are also reported in Pape *et al.* (2007).

2.5 Clustering

In this section, we present a clustering approach which yields an FBP for each cluster and which discards TFs which do not have any sufficient similarity. The clustering consists of three main steps:

(1) Selection: Select the pair with maximum similarity,
(2) Merging: Create the new FBP for the cluster,
(3) Verification: Discard the new cluster if not all members share sufficient similarity.

In the following, we describe each step in more detail. First, we have to change notation slightly. Let $\mathcal{Z} = \{Z_i\}$ be the set of TFs. The goal is to obtain a set $\mathcal{C}$ of disjoint classes $C_j \subseteq \mathcal{Z}$. The FBP/representative of class j is given by $r(C_j)$ while $r(\mathcal{C})$ obtains the set of all FBPs. Furthermore, $c(\cdot)$ returns the class index of a (meta) TF. We initialize the set of classes $\mathcal{C} = \mathcal{Z}$ to one class for each TF such that $c(Z_i) = i$ and $r(C_i) = Z_i$.

2.5.1 Selection This step selects pairs $X, Y \in r(\mathcal{C})$ which have highest similarity. Instead of using the introduced similarity measure directly, we create a related measure which automatically returns the overlap position with the maximum similarity. For each shift, we consider the ratio of the overlap probability and the probability of two independent hits of X and Y. Obviously, the denominator corresponds to the probability of two hits under a null model where X and Y are independent. In contrast, the numerator contains the probability of two hits considering the dependencies between X and Y. Applying the logarithm to the ratio yields log-odds scores:

$$S_{X,Y}(k) := \log\left(\frac{\gamma_{X,Y}(k)}{\alpha_X \cdot \alpha_Y}\right).$$

Taking the maximum over all shifts k and all pairs of X, X' and Y, Y', we can define the similarity measure $S^{\max}(X,Y)$:

$$\begin{aligned} S^{\max}(X, Y) := \max\Big(& \max_k S_{X,Y}(k), \max_k S_{X',Y}(k), \\ & \max_k S_{Y,X}(k), \max_k S_{Y',X}(k) \Big). \end{aligned}$$

Again, we are using certain equalities derived from the symmetric background model, in detail: $S_{X,Y}(k) = S_{X',Y'}(k)$, $S_{X',Y}(k) = S_{X,Y'}(k)$ and correspondingly for Y followed by X. We select X^*, Y^* such that

$$(X^*, Y^*) = \operatorname{argmax}_{(X,Y) \in r(\mathcal{C}) \times r(\mathcal{C}), X \neq Y} S^{\max}(X, Y).$$

Furthermore, we introduce a threshold d to stop clustering if the similarity is not sufficiently high. We set the parameter d equal to the 95% quantile of all pair-wise $S^{\max}$ values from the initial set $\mathcal{Z}$.

2.5.2 Merging After selecting one pair of TFs, we create the new FBP W: Let k^* denote the position of maximum overlap probability. The new FBP consists of the sum of the position count matrices of X and Y shifted by k^* positions. Thus, the length of W is $n_W = n_A + n_B - k^*$. If the maximum similarity occurs for X' or Y', we transform the respective position count matrices accordingly before summation. Before summation, we enlarge both position count matrices such that they overlap for each position of the FBP. The enlarged positions are filled with the background model. It is based on the background frequencies π_σ. Since we sum the position count matrices, we have to obtain counts. Therefore, we multiply π_σ by the average number of sequences of the corresponding position count matrix. This automatically corrects the information content of positions which do not overlap with all members of a cluster by adding the corresponding fraction of the background distribution. In other words, we take into account the number of motifs without specific signal at these positions.

2.5.3 Verification Due to the naive merging of TFs, the FBP might get less and less related to its original members after successive mergings. Hence, the clusters are no longer homogeneous but become more and more heterogeneous over the number of clustering steps. To prevent this, we ensure that the new FBP always has a high similarity to each of its members. The merge of the pair X and Y is discarded if any of the following inequalities does not hold:

$$\forall_{V \in C_{c(X)}, W \in C_{c(Y)}} \quad : \quad S^{\max}(V, W) > d$$

In case all inequalities hold, we update $\mathcal{C}$ by removing X and Y and adding the new cluster with its FBP W and members $C_{c(X)}$ and $C_{c(Y)}$. If at least one inequality does not hold, we skip the merging and mark the pair X and Y such that they cannot be merged. Then, the procedure starts with the selection step, again. The three steps are iterated until no non-marked pair of TFs has a similarity greater than d.

2.6 Alternative approaches

Since we compare the new approaches with existing alternative approaches, we give a brief review of those in this subsection. The alternative approaches presented here are based on a column-by-column comparison. The course of action is the same for all approaches: In the first step, a score or P-value is obtained for each position for each possible shift/gapless alignment. In a second step, the scores/P-values for each position are summarized yielding a score/P-value for each shift. Finally, the score/P-value for all shifts are summarized to one final value.

2.6.1 χ^2 test As introduced in Schones *et al.* (2005), the χ^2 statistic is used to compute the probability whether two columns are drawn from the same multinomial distribution. Let $n_{X\sigma}$ be the number of bases $\sigma \in \Sigma$ for PFM X. The marginal for PFM X is $n_X = \sum_{\sigma \in \Sigma} n_{X\sigma}$. The nucleotide marginals for two PFMs X and Y are denoted by $n_\sigma = n_{X\sigma} + n_{Y\sigma}$. The overall number of counts is n^*. Denoting the observed number of counts with an upper index o and the expected number of counts by $n_{X\sigma}^e = n_X \cdot n_\sigma / n^*$, we obtain a P-value using a χ^2 statistic with three degrees of freedom:

$$\sum_{\sigma \in \Sigma} \left(\frac{\left(n_{X\sigma}^o - n_{X\sigma}^e\right)^2}{n_{X\sigma}^e} + \frac{\left(n_{Y\sigma}^o - n_{Y\sigma}^e\right)^2}{n_{Y\sigma}^e} \right) \sim \chi_3^2$$

The P-values for all columns are summarized using the geometric mean. The final P-value is the minimum of the P-values for each shift.

2.6.2 Kullback-Leibler distance The Kullback-Leibler distance (Kullback, 1959) is often used as a similarity measure in this context (Aerts *et al.*, 2003; Roepcke *et al.*, 2005). Using above notation the symmetric form is defined by

$$\frac{1}{2}\left[\sum_{\sigma\in\Sigma}\left(\frac{n^o_{X\sigma}}{n^o_X}\log\frac{n^o_{X\sigma}\cdot n^o_Y}{n^o_X\cdot n^o_{Y\sigma}}+\frac{n^o_{Y\sigma}}{n^o_Y}\log\frac{n^o_{Y\sigma}\cdot n^o_X}{n^o_Y\cdot n^o_{X\sigma}}\right)\right]$$

The distances for the positions are summarized using the mean. The overall distance is obtained by taking the maximum over all shifts.

2.6.3 Tomtom using euclidean distance The Tomtom algorithm (Gupta *et al.*, 2007) can use any column-by-column similarity measure. The authors show that the euclidean distance introduced in this area by Choi *et al.* (2004) performs best. The distance is defined by

$$-\sqrt{\sum_{\sigma\in\Sigma}\left(\frac{n^o_{X\sigma}}{n^o_X}-\frac{n^o_{Y\sigma}}{n^o_Y}\right)^2}.$$

The sum of the distances for all position are the so-called raw scores. The Tomtom algorithm approximates a null distribution of these raw scores to obtain a *P*-value. The *P*-values for all n_k shifts are summarized by computing the *P*-value for the smallest observed *P*-value p^* by $1-(1-p^*)^{n_k}$.

2.7 Data

In this section, we describe the simulation, the Transfac and Jaspar set of PFMs and their preprocessing. In a first step, we compute PSSMs from the PFMs by taking the log-likelihood ratio of the nucleotide frequencies of the binding site and the background model. To ensure strictly positive ratios one adds pseudocounts in a step called regularization. We add pseudocounts to the position specific distributions according to the information content of the position (Rahmann, 2003). In fact, positions with low information content are shifted towards the background distribution. For positions with high information content, the difference to the background distributions is enforced. In general, one has to determine a threshold for each PFM. The threshold controls the probabilities of the type I error α and the type II error β given by $\alpha=\mathbb{P}_{H_0}(s\geq t)$ and $\beta=\mathbb{P}_{H_1}(s<t)$, where H_0 is the null model for random sequences and H_1 the model for the binding site. Probabilities α and β can be computed by the convolution of the position-specific scores and the respective nucleotide probabilities as weights (Rahmann, 2003). We set the threshold such that the probability of at least one false positive hit on a sequence of length 500 is $\leq 10\%$ and if possible $\alpha=\beta$ (see Pape *et al.* (2006) for details).

2.7.1 Simulation We compare the new similarity measures to existing approaches by using a simulation as reference: 10 000 sequences of length 10 000 are generated with an arbitrarily selected GC-content of 50%. After detecting all binding sites for a set of PFMs each with a threshold as defined earlier, we compute the number of overlapping hits N_{AB} between all pairs of TFs A and B. Based on these counts, we compute the simulated similarity as $\hat{S}(A,B)=N_{AB}/N_A$ where N_A denotes the number of hits of TF A. We get a symmetrical measure by using the average: $\hat{S}^{sym}(A,B)=(\hat{S}(A,B)+\hat{S}(B,A))/2$. In addition, we compare $S^{\max}$ with $\hat{S}^{\max}(A,B)=\max\{\hat{S}(A,B),\hat{S}(B,A)\}/2$.

The comparison is visualized by scatter plots for all pair-wise similarities. One dimension corresponds to the simulated similarity while the other dimension shows the computed similarity. We quantify the agreement between both measures using the Pearson correlation coefficient.

2.7.2 Sampling PFMs We generate a family of PFMs where the members are gradually more similar to each other. We sample the PFM column by column using a Dirichlet distribution with different parameter sets (Schones *et al.*, 2005). The blueprint is the consensus sequence 'ACGTACGT'. We choose this sequence because it contains palindromic as well as repeat features. Such features are crucial for a realistic test setting since they determine the overlap probabilities. The count matrix is based on 60 sequences where 30 sequences have the consensus letter at each position and 10 sequences for each of the other nucleotides. The counts for each position serve as parameters for the Dirichlet distribution to sample the multinomial frequency distribution per position. Thus, we have one parameter set for consensus letter 'A': (30, 10, 10, 10), one for 'C' (10, 30, 10, 10), and so on for 'G' and 'T'. To modify the sharpness of the Dirichlet distribution, we multiplied the parameter set by a power of ten for some PFMs (see Supplementary Material). Furthermore, we shift the PFM relative to the consensus. In combination, we also reduced the length or added positions with samples from a Dirichlet distribution with non-informative parameters (1, 1, 1, 1). In this manner, we sampled 10 PFMs, see Supplementary Material for sequence logos (Crooks *et al.*, 2004).

2.7.3 Transfac PFMs As a further test set, we used a vertebrate subset of Transfac (Matys *et al.*, 2003) PFMs of version 11.1. We selected 279 of the 588 vertebrate PFMs due to the following filtering: The position specific nucleotide distributions for some PFMs are similar to the background distribution. In these cases, they cannot be used for binding site detection since the score for a binding site is not significantly higher than a score for a random sequence. Such PFMs can be selected by assessing the average information content per position and the power of a PFM (Rahmann *et al.*, 2003). Thus, PFMs are discarded if either they have an average information content $<50\%$ or a type II error β based on the balanced threshold greater than 15%. Instead of using the balanced threshold for sequence annotation, we always set α to 10%. Otherwise, very powerful PFMs have such a small α that hits occur rarely and, therefore, the simulation obtains too few overlapping hits leading to bad estimates for the simulated similarity values.

2.7.4 Sequences The similarity for the Transfac PFMs is computed for two sets of sequences: random sequences and human promoter sequences. The random sequences are generated as above but with an average GC content equal to the human promoter sequences (44.86%). The human promoter sequences are based on Ensembl v46 (Hubbard *et al.*, 2005). For each Ensembl ID, we take the sequence region $-10\,000$ to $+200$ relative to the transcription start site. If this sequence overlaps with another Ensembl gene entry, we cut the sequence at that position.

2.7.5 Clustering of Jaspar PFMs The clustering is based on Jaspar (Sandelin *et al.*, 2004) PFMs using the data set analyzed by Mahony *et al.* (2007). The set consists of 13 classes each with closely related members. The classes are bZIP cEBP (4 members), bZIP CREB (4), bHLH (10), ETS (7), Forkhead (8), high mobility group (HMG: 6), HOMEO (8), MADS (5), NUCLEAR (8), REL (6), TRP (5), zinc finger DOF (4) and zinc finger GATA (4).

We assess the consistency of the clusters using the leave-one-out-cross-validation (LOOCV) approach following Sandelin *et al.* (2004) and Mahony *et al.* (2007): For each PFM except the singletons, we remove its contribution to the corresponding FBP. Then, we compute the similarity between the PFM and all FBPs and singletons. If the similarity between the PFM and its corresponding (modified) FBP is maximal, we call it a correct classification. A high percentage of correct classifications suggests a consistent clustering. In contrast to Mahony *et al.* (2007), we do not count singletons as misclassifications. Otherwise, more singletons in the clustering automatically lead to a lower consistency although more homogeneous clusters might have been retrieved. This occurs as soon as some PFMs only share weak similarity with all other PFMs. Hence, we

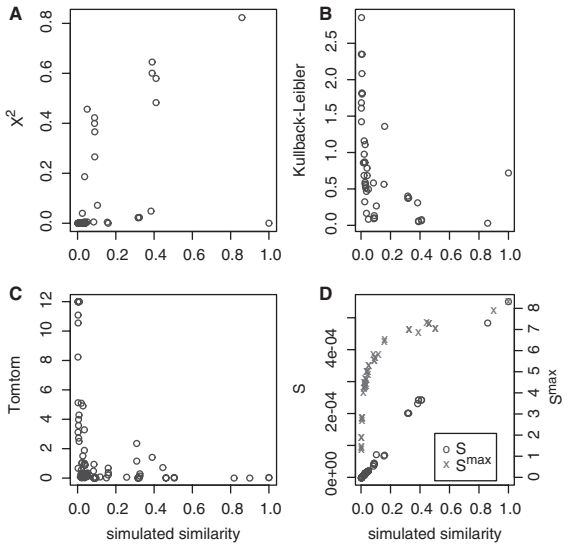

Fig. 1. Scatter Plot of all pair-wise similarities for the simulation (x axis) and the calculated similarity (y axis). (**A**) contains χ^2, (**B**) the Kullback-Leibler distance, (**C**) the Tomtom result using the euclidean distance, and (**D**) consists of the asymptotic covariance S (blue circles, left axis) and S^{max} (red crosses, right axis).

include singletons as FBPs for classifying although we do not classify the singletons.

3 RESULTS

3.1 Comparison with alternative approaches

In this article, we propose two new measures for similarity between PFMs. The first measure S is the asymptotic covariance between the number of hits of two TFs. For the purpose of clustering, we introduced the related measure S^{max} which computes the maximum log-odds score for the overlap probabilities. Figure 1 compares the new and three alternative measures with the measure computed by simulations. In Figure 1A the χ^2 test is compared with simulation. One observes a rough correlation although the highest simulated pair-wise similarity has a χ^2 similarity of 0. The Kullback-Leibler distance is shown in Figure 1B. Of course, the pairs with high Kullback-Leibler distance correspond to low simulated similarities. The measure can be used to separate similar and dissimilar pairs of TFs without too many false positives, e.g. with a low cut-off of 0.5. In addition, visualization of the rank transformed values shows that a rough ordering of similar pairs is possible (see Supplementary Material). The Tomtom approach based on the euclidean distance (see Fig. 1C) shows a similar behavior. In general, more pairs with high simulated similarity have a very small computed similarity. In contrast, the asymptotic covariance in Figure 1D denoted by S shows a strong linear correlation. There are no crucial disagreements between the simulation and the computation. The measure S^{max} which only captures the highest similarity grows with the simulated values but flattens for high values. Since we only consider the maximum overlap probability, differentiation between highly similar PFMs becomes more difficult. Still, an ordering is possible also for these values as shown in the rank transformed plots in the Supplementary Material.

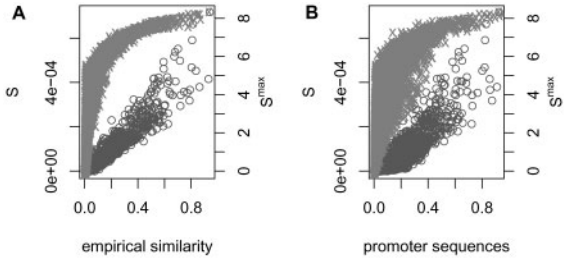

Fig. 2. Scatter Plot of each pair-wise similarity of Transfac PFMs between S (blue circles) and S^{max} (red crosses) based on (**A**) simulated sequences and (**B**) human promoter sequences.

A quantitive comparison value is given by the Pearson coefficient for the correlation between the simulated measure and the computed similarity measure. We obtain a Pearson coefficient of 0.509 (0.786 after rank transformation) for the χ^2 measure. The Kullback-Leibler distance, which is a distance instead of a similarity, has a negative correlation coefficient of −0.402 (−0.803). Although in this case, the Pearson correlation is a bad measure since the regression line is perpendicular to the x axis (see Fig. 1B). The distance from the Euclidean Tomtom approach has a small correlation coefficient of −0.292 (−0.674). The asymptotic covariance shows a strong linear correlation with a Pearson correlation coefficient of 0.997 (0.993). The measure S^{max} obtains a correlation coefficient of 0.76 (0.986).

3.2 Transfac set

Figure 2A shows the analysis on simulated sequences for the pairs of 279 Transfac PFMs. The asymptotic covariance has a strong linear correlation while, again, S^{max} flattens for higher similarity values. The analysis for human promoter sequences (Fig. 2) is similar but in general more scattered. The Pearson coefficient for S on simulated sequences is 0.952. The maximum measure obtains a Pearson coefficient of 0.615. The Pearson coefficient for the human promoter sequences for S is smaller (0.886) than for simulated sequences. In contrast, the Pearson coefficient increases for the maximum measure to 0.665. This is mainly due to the fact that the correlation is non-linear and the correlation coefficient is supported by the higher variance for low similarity values.

3.3 Clustering

The Jaspar set contains classes of closely related TFs. The clustering of the 13 classes with a total of 79 PFMs yields 14 clusters which are shown in the Supplementary Material. Three of four bZIP EBP PFMs are contained in the clustering, forming one homogeneous cluster. All four bZIP CREBs also form one homogeneous cluster. Eight of eleven bHLH are clustered forming two homogeneous clusters (size three and five). All seven ETS factors belong to one homogeneous cluster. All six Forkhead PFMs belong to one cluster which also contains four HMGs in a separate branch. One of the remaining two HMGs is clustered in a small cluster with one HOMEO. Five of the remaining seven HOMEOs are in one

homogeneous cluster, as well, as all five MADS PFMs. Seven of eight NUCLEAR receptors are contained in one homogeneous cluster. All six RELs belong to one homogeneous cluster. Two of the five TRPs are contained in a heterogeneous cluster with all four zinc finger DOFs, two of the remaining three TRPs are also clustered together homogeneously. Finally, two of the four zinc finger GATAs are forming one homogeneous cluster. Altogether, eleven of the 14 clusters are homogeneous, containing 49 of 67 PFMs while twelve PFMs are not clustered at all.

We compare our clusters (including the eight zinc fingers) with the corresponding results from the clustering in Mahony *et al.* (2007) based on an ungapped Smith-Waterman alignment with the Pearson correlation coefficient as scoring function. This clustering from Mahony *et al.* (2007) including the eight zinc fingers is very similar to the clustering without the zinc fingers in Figure 8 (Mahony *et al.*, 2007) but yields 16 clusters and two singletons (personal communication). The subtle differences are considered below. We yield seven times the same clusters: ETS, Nuclear Receptor, bZIP CREB Subgroup, bZIP cEBP Subgroup, MADS, HOMEO Subgroup and the TRP-cluster/IRF Subgroup with zinc finger DOFs. Another five clusters are modified: The REL-like group becomes a homogeneous cluster since En1 from the HOMEO group and Chop-cEBP from the bZIP group are removed. The bHLH-ZIP cluster does not contain the correct member Arnt-Ahr any more. The TRP-cluster/Myb Subgroup lacks the correct member MYB-ph3. The HMG/Forkhead Group 1 does not contain the wrong member Pbx from the HOMEO group. Our HMG/HOMEO mix contains different TFs from the same classes HMG and HOMEO, specifically HMG-1 and En1. We also obtain a cluster for the zinc finger GATA PFMs but only containing two in comparison to three in Mahony *et al.* (2007). Instead of the mixed cluster including the zinc finger GATA1, the bHLH TAL1-TCF3 and the Forkhead FOXL1, we obtain an extended bHLH Subgroup cluster with TAL1-TCF3 and the two other bHLHs NHLH1 and MYF. The two latter ones are clustered by Mahony *et al.* (2007) into a single cluster. In fact, all three members share the consensus CA*CTG justifying the extension of the cluster. The heterogeneous cluster HMG/Forkhead Group 2 with two members does not appear in our analysis.

Furthermore, we performed the LOOCV test on our clustering. All 67 PFMs are classified correctly while excluding the 12 singletons. The high number of correct classifications is not surprising since the clustering algorithm intrinsically computes a consistent clustering by testing in each step the similarity between all members and their respective FBP. In comparison, Mahony *et al.* (2007) obtain 72 of 77 correct classifications likewise without counting the two singletons as wrong classifications. Hence, Mahony *et al.* (2007) assign more PFMs to clusters while increasing the number of heterogeneous clusters and decreasing the consistency of the clustering. Instead, we obtain more singletons leading to a more stringent and more consistent clustering.

The clustering automatically generates an FBP for each cluster. All FBPs are given in the Supplementary Material. As an example, the FBP for the bZIP CREBs is printed here in Figure 3. Since we automatically consider the number of supporting PFMs per position by using the background model

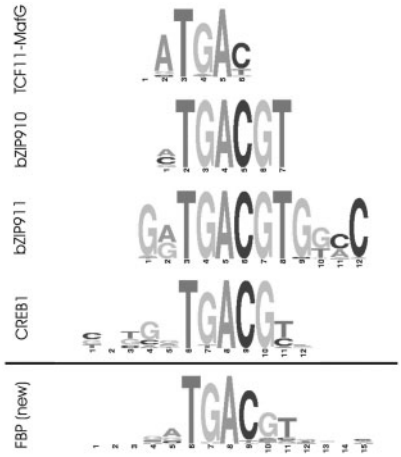

Fig. 3. FBP of the bZIP CREB cluster with the multiple alignment containing the four TFs TCF11-MafG, bZIP910, bZIP911 and CREB1.

for non-overlapping positions in each merging step, the corresponding positions do not obtain high information content.

4 DISCUSSION

We have introduced two new measures of similarity for PFMs. One main difference is that these new similarity measures depend on the regularization method, the parameter which represents the threshold to detect a hit, and on the background model. To remove redundancies in a set of PFMs, we consider this an advantage because the detected binding sites in a sequence also depend on these parameters. In fact, the similarity measure is able to capture the differences between the results of two different parameter sets. As an extreme example consider two different PFMs with the same length and both with a threshold such that all words are accepted. Due to their co-occurring hits, these PFMs have the highest similarity. Increasing the threshold decreases the number of words with scores higher than the threshold. Therefore, similarity should decrease as it does in the new approaches. The background model also can have a high influence on the results. For example, two overlapping PFMs without CpG dinucleotides are very similar within a CpG island because both differ from the background. In a CpG poor background model, both PFMs are hidden within the background, thus, neither they achieve a high similarity nor are their hits correlated. Again, this is advantageous for removing redundant PFMs in a set. Furthermore, extending the similarity measure for higher order Markov models is possible although calculation of the 2D score distribution will become time consuming.

In contrast to the advantageous effect on the removal of redundancies, the dependence of the result on the parameter choices is unwanted for clustering. Although the clustering is robust against small changes, of course, big differences in the parameters do change the result. For example, changing the GC content to 40% changes the composition of four clusters by 1–2 insertions/deletions per cluster and adds a new small cluster of size two. Instead, substituting the regularization method by a

simple addition of pseudocounts (0.01) only has a minor effect by changing the composition of one cluster slightly.

The analysis considering 279 Transfac PFMs shows that the similarity measure is not only applicable to artificial PFMs but also to real binding sites. The comparison between simulated sequences and human promoter sequences shows that for no pair of TFs the simulated similarity significantly differs from the theoretical similarity. Large differences, e.g. high simulated similarity and low theoretical similarity, might give evidence for competitive binding due to more overlapping binding sites than expected by chance. Since we do not observe such deviations, either signal to noise ratio is too low or competitive binding sites evolve to be similar regarding their sequence.

The clustering of the Jaspar set yields a high number of homogeneous clusters. In addition, only a minor fraction of PFMs are not clustered at all. Hence, it seems that, indeed, the similarity between PFMs is captured appropriately. Furthermore, the clustering yields a FBP/representative for each cluster containing the characteristic properties of its class members.

In this article, we have introduced two new measures of similarity. In contrast to existing measures, we give a natural interpretation of the similarity, which is especially useful in practice. We use a statistical framework to derive the measure, resulting in the asymptotic covariance. To the best of our knowledge, we give the first formulas and efficient calculation in the context of PFMs. Of course, the measures can also be applied to arbitrary set of words, i.e. experimentally verified binding sites, although computation becomes inefficient for large sets. We also show that the similarity measure outperforms existing approaches and that the concept can successfully be applied to clustering.

ACKNOWLEDGEMENTS

Thanks to Hugues Richard for fruitful discussions and Holger Klein for a program to count the overlapping hits between pairs of TFs. We also thank Shaun Mahony for making the JASPAR set and additional STAMP analyses available for us. The anonymous reviewers gave valuable comments which improved the manuscript significantly.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al*. (2003) Computational detection of cis -regulatory modules. *Bioinformatics*, **19** (Suppl. 2), ii5–ii14.

Bailey,N.T.J. (1977) *Mathematics, Statistics and Systems for Health*. Wiley, NY.

Beckstette,M. *et al*. (2006) Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, **7**, 389.

Berg,O. and von Hippel,P. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Choi,I.-G. *et al*. (2004) Local feature frequency profile: A method to measure structural similarity in proteins. *PNAS*, **101**, 3797–3802.

Claverie,J.-M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.

Crooks,G.E. *et al*. (2004) Weblogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Fleiss,J.L. *et al*. (2003) *Statistical Methods for Rates and Proportions*. John Wiley & Sons, NY.

Gupta,S. *et al*. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

Hertz,G. *et al*. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.

Kielbasa,S.M. *et al*. (2005) Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, **6**, 237.

Kullback,S. (1959) *Information Theory and Statistics*. John Wiley & Sons, New York, USA.

Liu,J.S. *et al*. (1990) Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *J. Am. Stat. Assoc.*, **95**.

Mahony,S. *et al*. (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, **21** (Suppl. 1), i283–i291.

Mahony,S. *et al*. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.

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

Narlikar,L. and Hartemink,A.J. (2006) Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, **22**, 157–163.

Pape,U.J. *et al*. (2006) A new statistical model to select target sequences bound by transcription factors. *Genome Informatics*, **17**, 134–140.

Pape,U.J. *et al*. (2007) Compound Poisson approximation of DNA motif counts on both strands. Accepted by J. Comput. Biol.

Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments published erratum appears in nucleic acids res 1996 nov 1;24(21):4372. *Nucleic Acids Res.*, **24**, 3836–3845.

Rahmann,S. (2003) Dynamic programming algorithms for two statistical problems in computational biology. In *Proceedings of the 3rd Workshop of Algorithms in Bioinformatics (WABI)*. Springer Verlag, Heidelberg, pp. 151–164.

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

Reinert,G. *et al*. (2000) Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, **7**, 1–46.

Roepcke,S. *et al*. (2005) T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res.*, **33** (Suppl. 2), W438–W441.

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

Sandelin,A. *et al*. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.

Schneider,T. *et al*. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Schones,D.E. *et al*. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12** (1 Pt 2), 505–519.

Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.*, **5**, 89–96.

Stormo,G. and Hartzell,G. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. *et al*. (1982) Use of the ''Perceptron'' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3012.

Suzuki,M. and Yagi,N. (1994) DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.

Wasserman,W.W. and Sandelin,A. (2004) Applied Bioinformatics for the Identification of Regulatory Elements. *Nat. Rev. Genet.*, **5**, 276–287.

Waterman,M.S. (2000) *Introduction to Computational Biology*. chapter 12: Probability and Statistics for Sequence Patterns, Chapman & Hall/CRC.

Wu,T.D. *et al*. (2000) Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**, 233–244.

Zhang,J. *et al*. (2007) Computing exact P-values for DNA motifs. *Bioinformatics*, **23**, 531–537.