BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 9 2010, pages 1152–1159
doi:10.1093/bioinformatics/btq106

Sequence analysis

Advance Access publication March 11, 2010

Localized motif discovery in gene regulatory sequences

Vipin Narang[1], Ankush Mittal[2] and Wing-Kin Sung[1,*]

[1]Department of Computer Science, National University of Singapore, Singapore – 117417 and [2]Department of Computer Science and Engineering, College of Engineering Roorkee, Uttarakhand – 247667, India

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Discovery of nucleotide motifs that are localized with respect to a certain biological landmark is important in several applications, such as in regulatory sequences flanking the transcription start site, in the neighborhood of known transcription factor binding sites, and in transcription factor binding regions discovered by massively parallel sequencing (ChIP-Seq).

Results: We report an algorithm called LocalMotif to discover such localized motifs. The algorithm is based on a novel scoring function, called spatial confinement score, which can determine the exact interval of localization of a motif. This score is combined with other existing scoring measures including over-representation and relative entropy to determine the overall prominence of the motif. The approach successfully discovers biologically relevant motifs and their intervals of localization in scenarios where the motifs cannot be discovered by general motif finding tools. It is especially useful for discovering multiple co-localized motifs in a set of regulatory sequences, such as those identified by ChIP-Seq.

Availability and Implementation: The LocalMotif software is available at http://www.comp.nus.edu.sg/~bioinfo/LocalMotif

Contact: ksung@comp.nus.edu.sg

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

Received on May 20, 2009; revised on February 19, 2010; accepted on March 4, 2010

1 INTRODUCTION

Regulation of gene expression primarily occurs at the stage of transcription of the gene. A key event in the regulation of transcription is the binding of *trans*-acting proteins called transcription factors (TFs) to *cis*-acting DNA sequences in the vicinity of the gene. The TFs bind to short 5–20 bp segments of DNA called transcription factor binding sites (TFBSs). The TFBSs are not easily recognized in DNA sequences as they are of short length and frequently have mutations in their pattern. However, they can be discovered computationally within a set of sequences that are enriched with their occurrences. The computational algorithm looks for a short, conserved and often repeated pattern called the *motif* in these sequences. The motif is likely to be the TFBS. A number of computational algorithms for motif discovery have been developed over more than a decade (Bailey and Elkan, 1994; Buhler and Tompa, 2002; Eskin and Pevzner, 2002; Ettwiller *et al.*, 2007; Fratkin *et al.*, 2006; Henikoff *et al.*, 1995; Lawrence *et al.*, 1993; Linhart *et al.*, 2008; Liu *et al.*, 2001; Marsan and Sagot, 2000; Pavesi *et al.*, 2004; Pevzner and Sze, 2000; Roth *et al.*, 1998; Thijs *et al.*, 2002).

Motif finding algorithms can usually discover the real motif (i.e. which truly represents the TFBSs) when the TFBSs are significantly over-represented compared to random (or background) patterns in the given set of sequences. However, when analyzing long sequences or a set of sequences in which the TFBSs are not significantly enriched, random patterns can appear equally or more conserved than the real motif and thus the real motif cannot be discovered (Buhler and Tompa, 2002; Keich and Pevzner, 2002a, 2002b). In such datasets, additional information about the real motif can aid its discovery.

A useful piece of information that has not been adequately exploited in the existing motif finding algorithms is the positional localization of TFBSs. TFBSs usually occur in specific positions relative to a biological landmark within gene regulatory sequences. For instance, many TFBSs are located in specific position intervals relative to the transcription start site (TSS) (Smale and Kadonaga, 2003). TFBSs of cooperating TFs also occur at specific distances from each other (Vardhanabhuti *et al.*, 2007). Similarly, in a set of TF binding sequences obtained by high-throughput techniques such as ChIP-Chip or massively parallel sequencing (ChIP-Seq), the TFBSs are localized around the positions of maximum signal intensity (such as a peak in ChIP-Seq). Such information of positional localization of TFBSs can be utilized to distinguish the real motif from random patterns.

Localization information has been used previously to improve the performance of motif discovery. For example, Ohler *et al.* (2002) analyzed motifs in 1941 Drosophila regulatory sequences of length 300 bp each aligned (−250, +50) relative to the TSS. The analysis of complete 300 bp sequences did not reveal many of the core promoter motifs. However, in a separate analysis of the local region (−60, +40), most core promoter motifs were discovered. Similarly, Molina and Grotewold (2005) analyzed the (−50, −1) and (+1, +50) regions of *Arabidopsis thaliana* promoters separately in order to discover the core promoter motifs. Vardhanabhuti *et al.* (2007) considered both positional localization around the TSS and pairwise distances among the motifs to identify novel TFBSs in human promoters. Qi *et al.* (2006) imposed a positional prior on the motif in TFBSs identified by ChIP-Chip to improve motif discovery. The prior was proportional to the ChIP signal intensity at a given position, which was estimated by combining the intensities of all neighboring probes.

Recently, some generic motif finding algorithms have incorporated the facility to define positional priors during motif search. For example, the AMADEUS algorithm (Linhart *et al.*, 2008) incorporates the facility to score motifs according

*To whom correspondence should be addressed.

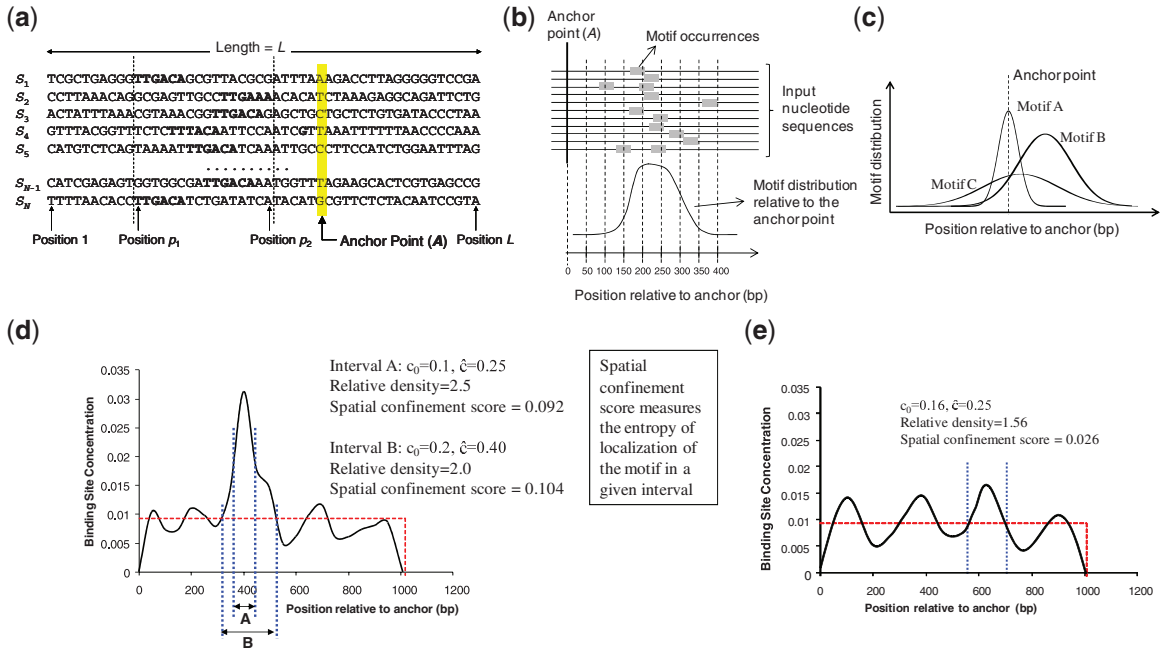

Fig. 1. (**a**) In the localized motif finding problem, the instances of a motif M occur confined within an unknown position interval (p_1, p_2) of the input sequences relative to an anchor point. The objective is to discover the motif M and the interval (p_1, p_2). (**b**) LocalMotif studies the positional distribution of the motif occurrences relative to the anchor point. (**c**) Different motifs could be localized within different intervals around the anchor point. (**d**) A *localized* motif is prominently over-represented in a certain interval with respect to the entire sequence length, (**e**) whereas a random pattern may be over-represented in certain intervals but not with respect to the entire sequence.

to their distribution around the TSS. Tharakaraman *et al.* (2005) incorporated positional preference in their motif finding algorithm GLAM by performing gapless local alignment over windowed subsequences of the original sequence set (aligned relative to the TSS) instead of the complete length. These applications use a positional prior on the motif, i.e. the localization interval of the motif is defined a priori. However, in most practical scenarios both the position and the length of the localization interval are unknown. None of the existing algorithms has addressed this problem.

A solution could be to subdivide the sequences (aligned relative to the biological landmark) into short intervals and analyze each interval separately with the existing algorithms. However, this is impractical due to several reasons. First, it is difficult to decide the interval length. If the interval is too short or too long compared to the actual region of localization, the motif will not be discovered. Furthermore, discovering multiple motifs spread over different intervals would require selection of different interval lengths (Fig. 1c). Secondly, this approach would report a number of random patterns that are over-represented in a short sequence interval by chance, but which are not truly localized motifs. There is a difference between a *localized motif* and a motif that is over-represented in a short sequence interval. As shown in Figure 1d and e, a *localized motif* has a distinct confinement of TFBSs in a certain interval in the context of the entire sequence length, while a motif that is over-represented in a certain short interval may not have such confinement in the global context. Therefore, local analysis without considering the global context of the motif may be misleading. This is illustrated in the study of Friberg *et al.* (2005), where scoring functions with a positional bias resulted in a large number of false motifs. Thirdly, the task of fragmenting the sequences and combining together results for several intervals is laborious and time consuming. Thus, it would be useful to have an automated, efficient algorithm to accurately discover localized motifs.

This article presents a computational algorithm called LocalMotif for the discovery of localized motifs in sequences that have been aligned relative to a biological landmark (henceforth referred to as the *anchor point*). A new scoring measure called *spatial confinement score* is introduced that assesses whether or not a motif has localized occurrence within the sequences, and allows accurate demarcation of the localization interval. The spatial confinement score is combined with the existing scoring measures of motif over-representation and relative entropy to evaluate the overall prominence of the motif. The existing scoring measures are reformulated using information theory so that all scores can be easily combined into a single score. A time and memory efficient greedy search algorithm utilizes this scoring function for localized motif discovery.

Experiments on simulated datasets show that LocalMotif has consistently better performance than existing tools in detecting motifs in cases where they are localized in a certain sequence interval. The interval length predictions made by LocalMotif are also highly accurate. Experiments on real datasets show that LocalMotif can discover localized motifs around the TSS, co-regulatory motifs around known motifs and motifs in ChIP-Seq datasets where the existing motif finding tools fail. Furthermore, the interval predictions made by LocalMotif provide biologically useful information about TF–TF interactions.

2 METHODS

The LocalMotif algorithm is described below in the following aspects: the *motif model*, the *scoring function* and the *algorithm*.

2.1 Motif model

The LocalMotif algorithm has two modules—a core module that discovers prominent non-redundant motifs, and a refinement module that fine-tunes these motifs.

The core module uses the consensus (l, d) representation, which describes the motif as a nucleotide pattern of length l such that any binding site differs from this pattern up to a maximum of d point substitutions (Pevzner and Sze, 2000). This representation allows mismatches to occur at any position within the motif with equal frequency. Although this is not a valid assumption in real protein–DNA interactions where some positions are more admissive of base substitutions than the others, it is advantageous for *ab initio* motif finding as it reduces the search space and does not impose an initial assumption on the nature of mutations in the motif (Keich and Pevzner, 2002a; Pavesi *et al.*, 2004; Pevzner and Sze, 2000; Sinha and Tompa, 2003).

The motifs discovered by the core module are provided to a refinement module. The refinement module uses the positional weight matrix (PWM) representation (Stormo, 2000), which more accurately represents the different binding preferences at various positions in a motif. For each (l, d) motif discovered by the core module, the refinement module determines the optimal PWM.

2.2 Problem formulation

The localized motif finding problem is stated here as a modification of the (l, d) motif problem defined by Pevzner and Sze (Pevzner and Sze, 2000). Consider a set of N input DNA sequences $\mathbf{S}=\{S_1, S_2, \ldots, S_N\}$ of length L each, aligned relative to an anchor point A (which is a biological landmark such as the TSS) as shown in Figure 1a. Suppose that the instances of an unknown (l, d) motif M occur confined within an unknown interval (p_1, p_2) of the sequences. The objective is to discover M and (p_1, p_2) given $\mathbf{S}$, l and d.

2.3 Scoring function

LocalMotif combines three different scoring functions that individually describe three different characteristics of a motif: the relative entropy score (RES) which measures the degree of surprise in the motif nucleotide pattern with respect to the background distribution of nucleotides, the over-representation score (ORS) which measures the overabundance of the number of instances of the motif relative to background, and the spatial confinement score (SCS) which measures the disproportionate confinement of motif instances in a certain sequence interval. While the former two scoring measures exist in the literature, the spatial confinement score has been introduced in LocalMotif to aid the discovery of localized motifs. All scoring measures are brought to a consistent form as normalized entropy measurements, so that they may be combined together and are comparable across motifs with different (l, d). Detailed derivations of the formulae are provided in Supplementary Section A. Computation of RES and ORS requires definition of a background model for the sequences. Following the state-of-the-art motif finding algorithms, LocalMotif uses an order-q Markov process to model the background (Thijs *et al.*, 2002), where q is a user-defined constant.

2.3.1 Over-representation score A motif is enriched in the sequences if its number of observed instances significantly exceeds the number of instances expected by chance (according to background). The ORS is a statistical measure of the difference between the observed and chance occurrence counts. In random sequences sampled from a Markov background, the number of chance occurrences follows the binomial distribution. Let e_0 be the chance proportion of the instances of motif M in the background. Among any t observed patterns, the probability of observing k instances is given by the binomial formula $P(k,t|e_0) = {}^nC_k (e_0)^k (1-e_0)^{t-k}$. Now let the number of instances of the motif M in the sequences be a proportion, e_1, of all $N(L-l+1)$ length l patterns in the sequences. The ORS of M is measured as the Kullback–Leibler divergence between the binomial distributions $P(k,t|e_0)$ and $P(k,t|e_1)$:

$$\mathrm{ORS} = D(E_0||E_1) = \frac{1}{\phi(l,d)}\left[e_0 \ln\left(\frac{e_0}{e_1}\right) + (1-e_0)\ln\left(\frac{1-e_0}{1-e_1}\right)\right], \quad (1)$$

where $\phi(l,d) = (1/4^l)\sum_{i=0}^{d} {}^lC_i 3^i$ is a normalization factor equal to the fraction of length l patterns that have up to d mismatches from a given pattern.

2.3.2 Relative entropy score The TFBSs are expected to be distinct from the background since the TF can distinguish them from surrounding nucleotide patterns. RES (Hertz and Stormo, 1999; Stormo, 2000; Thijs *et al.*, 2002) measures the difference between the motif and background. Let all observed TFBSs of the motif be aligned vertically, and the average frequency of occurrence of each nucleotide $b \in \{A, C, G, T\}$ at each position $i = 1, 2, \ldots, l$ be $f_{b,i}$. The entropy of the motif M relative to the background model B is usually measured as the Kullback–Leibler divergence $D(M \| B)$:

$$\mathrm{RES} = D(M||B) = \sum_{i=1}^{L}\sum_{b} f_{b,i} \ln\left(\frac{f_{b,i}}{p_b}\right) \quad (2)$$

where $p_b, b \in \{A, C, G, T\}$ are the *a priori* frequencies of the nucleotides in the background. The RES is normalized as:

$$\mathrm{RES}_{\mathrm{norm}} = \frac{1}{l \ln 4}\sum_{i=1}^{l}\sum_{b} f_{b,i}\ln(f_{b,i}) - \frac{1}{\ln 4}\sum_{b}\bar{f}_b \ln(p_b) \quad (3)$$

where $\bar{f}_b = \frac{1}{l}\sum_{i=1}^{l} f_{b,i}$. The normalized RES is independent of the motif length l and usually lies in the range (0,1).

2.3.3 Spatial confinement score Motif finding algorithms usually consider the TFBSs to be randomly distributed across the entire sequence length. However, LocalMotif considers the non-uniform distribution of TFBSs in the sequences relative to the anchor point. Let c denote the proportion of TFBSs that fall within the interval (p_1, p_2), i.e. if n is the total number of TFBS across entire sequence length L, and n_1 is the number of TFBS in the interval (p_1, p_2), then $c = n_1/n$. If the TFBSs are uniformly distributed across the entire sequence length L, then it is expected that the proportion of TFBSs falling within any interval (p_1, p_2) will be $c = c_0 = |p_2 - p_1|/L$. For example, in any interval of length $L/2$ one would expect to find 50% of the TFBSs. However, if the TFBS distribution is non-uniform, the proportion would be higher in some intervals and lower in others. LocalMotif intends to discover the shortest interval that encompasses the maximum proportion of TFBSs. It thus compares the proportion of TFBSs that lies within the interval and the proportion that lies outside it. The interval that maximally separates the two has the highest *spatial confinement score*. Let $\hat{c}$ be the observed proportion of TFBSs that lie within an interval (p_1, p_2). The spatial confinement score for the interval is given by the entropy difference (KL-divergence) between the observed proportion, $\hat{c}$, and uniform proportion, c_0:

$$\mathrm{SCS} = D(\hat{c}||c_0) = \hat{c}\ln\left(\frac{\hat{c}}{c_0}\right) + (1-\hat{c})\ln\left(\frac{1-\hat{c}}{1-c_0}\right) \quad (4)$$

Note that a short interval with high density of TFBSs may not have a spatial confinement score as high as a longer interval with lesser density of TFBS if the longer interval encompasses a large proportion of the TFBS compared to its surroundings. For example, in Figure 1d the score for interval B is higher than that for interval A. Therefore, the maximization of SCS of a motif over all intervals gives the interval where the motif is maximally localized. In addition, a localized motif has high SCS since most of its TFBS are confined within the localization interval, whereas a locally over-represented motif has low SCS (Fig. 1e) since a significant fraction of its TFBS are also distributed in rest of the sequence length.

2.3.4 Combined score The three scoring measures mentioned above, *viz.* RES, ORS and SCS, measure three independent characteristics of a motif. They have been expressed as entropies measured as KL divergence between an observed and a reference probability distribution. Thus, they are independent of situational parameters such as motif length l, number of allowed substitutions d, sequence length L and the interval length $|p_2 - p_1|$. Furthermore, they have been normalized to usually range between (0, 1) and have consistent values barring extreme situations. The scores can thus be combined in various ways. One of the ways is to consider each score as a separate coordinate and define the total score as the distance from the origin. In this case, the combined score is:

$$\text{Linear combination score} = \sqrt{w_1 \text{RES}_{\text{norm}}^2 + w_2 \text{ORS}^2 + w_3 \text{SCS}^2}, \quad (5)$$

where w_1, w_2 and w_3 are user-specified weights for the three scores. By default, all the weights are set to 1. This linear combination score assigns higher rank to a motif that is exceptional in any one of the three scores. Another way could be a geometric combination of the three scores as:

$$\text{Geometric combination score} = |\text{RES}_{\text{norm}}|^{w_1} \cdot |\text{ORS}|^{w_2} \cdot |\text{SCS}|^{w_3}. \quad (6)$$

This score assigns higher rank to a motif that performs well in all three individual scoring measures. There could be various other ways of combining the scores. It remains an open problem how to optimally combine the three scores.

2.3.5 P-values The *P*-values of RES, ORS and SCS are computed individually. Since the KL divergence, D, is directly related to the likelihood ratio (LR) test (Eguchi and Copas, 2006), the Wilks' theorem can be used to estimate the *P*-value of D. The LR test statistic $\Lambda_{\text{RES}} = 2n \times \text{RES}$ is χ^2 distributed with $3l$ degrees of freedom, while the statistics $\Lambda_{\text{ORS}} = 2t(e_0 + e_1) \times \text{ORS}$ and $\Lambda_{\text{SCS}} = 2n \times \text{SCS}$ are both χ^2 distributed with one degree of freedom. The *P*-value can be computed as area under the tail of the χ^2 distribution to the right of the LR test statistic.

2.4 Algorithm

The core and refinement modules of the LocalMotif algorithm are briefly described below, with details provided in Supplementary Section B.

The LocalMotif core module scores candidate (l, d) motifs in different sequence intervals and reports the best scoring ones. An exhaustive enumeration strategy would require scoring all possible 4^l candidate patterns in all possible sequence intervals, leading to a complexity of $O(4^l.L^2)$. Therefore, a greedy search approach is used, where initially only the l-mers occurring directly within the sequences are considered as candidates. Scoring each candidate l-mer in all possible position intervals $(p_1, p_2) : 0 \le p_1 < p_2 \le L$, would be formidable. Thus, only the intervals $(p_1, p_2) : p_1 < p_2; p_1, p_2 \in \{0, s, 2s, 3s, \ldots, L\}$ are considered, where s, called step size, is a small integer value. Interestingly, the score for a longer interval can be computed directly from the scores for shorter constituent intervals, resulting in considerable computational savings (Supplementary Material). As the candidate l-mers are being scored in different position intervals, a list of top n scores is maintained, where n can be set depending upon available memory. If two candidates have similar pattern (similarity >65% evaluated using Needleman–Wunsch global alignment) and overlapping position intervals, the lower scoring candidate is discarded. Following this initial search, a heuristic algorithm similar to SP-STAR (Pevzner and Sze, 2000) extends the search to other probable patterns that do not occur directly within the sequences.

The results of separate runs with varying (l, d) are combined directly since the LocalMotif scoring function does not depend upon l and d. Between two motifs with similar pattern (similarity >65% measured relative to the shorter motif) and overlapping intervals, the one with lower score is discarded.

The (l, d) motifs discovered are then fed to the refinement module that generates an optimal PWM corresponding to each motif. The refinement module begins with an initial PWM for the motif constructed from all of its d-mismatch instances. The PWM is then updated iteratively by a Fitness Expectation Maximization (FEM) algorithm (Wierstra *et al.*, 2008), which seeks to maximize the LocalMotif scoring function for the PWM. A low value of forget factor is employed in the EM iterations so that the algorithm converges to a local minimum nearby the initial PWM in the solution space rather than approaching the global minimum. The algorithm converges within a few (<10) iterations giving the optimal PWM for a motif.

2.5 Implementation

The basic LocalMotif algorithm is implemented in platform independent C++, and is supplemented by a user-friendly interface written in Python. The source code and complied binaries are available freely at the authors' website: http://www.comp.nus.edu.sg/~bioinfo/LocalMotif. The user can specify the following parameters to suit the dataset and available computing resources: (i) Background model, (ii) Number of candidates n to be retained in memory, (iii) Maximum interval length (when analyzing long sequences, setting a maximum interval length such as 1 kb makes the analysis faster), (iv) Number of motifs to output, (v) Choice of single or double strand analysis, and (vi) choice of linear or geometric combination of the three scoring functions and their respective weights.

The program outputs the discovered motifs with their intervals of localization, the three individual scores (RES, ORS and SCS), and the combined (weighted) scores. The individual scores reveal the prominent characteristics of a motif and may be used to reject outliers.

3 RESULTS

The advantage of LocalMotif scoring function is first demonstrated in this section through a simulation experiment. Then the performance of motif discovery is evaluated over both synthetic and real datasets. Comparison is made with four other freely available motif finding tools: MEME (Bailey and Elkan, 1994), Weeder (Pavesi *et al.*, 2004), Trawler (Ettwiller *et al.*, 2007) and Amadeus (Linhart *et al.*, 2008).

3.1 Analysis of the scoring function

The LocalMotif scoring function is illustrated through a planted motif problem. Fifty sequences, each 3000 bp long, were generated using a zero-order uniform Markov model. An instance of a (7,1) motif ATGCATG was implanted in 75% of the sequences within the interval (2000, 2500). Mathematical analysis (Buhler and Tompa, 2002; Keich and Pevzner, 2002b) shows that the (7,1) motif is indistinguishable from random patterns within the 3000 bp length sequences, but is significantly over-represented in the 500 bp interval.

Five top scoring motifs reported by LocalMotif and their scores are shown in Table 1. The planted (7,1) pattern was correctly identified as the top motif and its interval of localization was accurately determined. Although the planted motif has a low ORS compared to competing random patterns, it has a substantially higher SCS of 0.485 as compared to the spurious motifs whose SCS is <0.3. The RES is similar for all motifs because of the uniform background. Thus, the planted motif is correctly recognized due to the SCS.

Contours of ORS and SCS for the planted motif in various position intervals are shown in Figure 2. The x- and y-axes in the contour plot correspond to the positions p_1 and p_2, respectively for the interval (p_1, p_2) and the shade of the contour indicates the magnitude of ORS or SCS in the interval. The ORS is large wherever there is a local concentration of binding sites, whereas SCS is large only in the motif's localization interval (2000, 2500). The SCS thus plays an important role in predicting the accurate interval of localization.

Table 1. Results of analyzing with Local Motif a set of simulated sequences of length 3000 bp containing a planted (7, 1) pattern ATGCATG

Motif Pattern	Motif interval	Motif score	Score components		
			RES	ORS	SCS
ATGCATG	(2060, 2445)	0.757	0.481	0.326	0.485
GGACGCT	(15, 115)	0.733	0.481	0.500	0.235
AGCGCCG	(455, 575)	0.712	0.481	0.439	0.289
GTCCGAT	(85, 200)	0.691	0.482	0.408	0.282
TCCCTGC	(2340, 2450)	0.690	0.481	0.411	0.275

Five top scoring motifs and their reported localization intervals are shown.

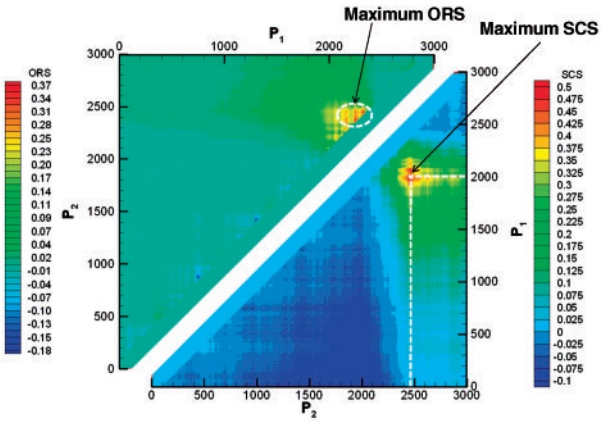

Fig. 2. Contours of local over-representation score (upper triangle), and spatial confinement score (lower triangle) of the planted motif ATGCGTAC in different position intervals (p_1, p_2) of the simulated sequence set.

Table 2. Ranges of parameters studied in simulated sequence data

Parameter	N	L	k	$\bar{p}$
Range	50–100	200–1000 bp	20–100%	0.1–1.0

3.2 Performance on simulated datasets

The test on simulated sequences evaluates the accuracy and robustness of motif and localization interval predictions made by Localmotif. Each dataset contains N nucleotide sequences of the same length L selected randomly from the human genome. In about k percentage of the sequences, a known binding site for a single TF obtained from TRANSFAC (Matys *et al.*, 2003) is implanted within the position interval $I=(p_1, p_2)$. A total of 100 such datasets were generated while randomly varying the parameters N, L, k, $\bar{p}$ and the TF as shown in Table 2. Note that $\bar{p}=|I|/L$ denotes the ratio of interval length to sequence length, with $|I|=(p_2-p_1)$. The TFs were chosen among 10 different vertebrate TFs each of which has at least 60 binding sites in the TRANSFAC database (refer Supplementary Section C). Thus, a fair variety of test conditions were simulated.

The performance of motif detection is shown in Figure 3. Accuracy of motif detection is measured according to whether the known motif is reported as the top scoring motif. Linear score combination with equal weights for the three scores was specified for LocalMotif. The same background model was used for all tools tested, which was constructed from the set of 1 kb upstream promoter and 1 kb exon sequences of all human refseq genes (Supplementary Section C). The motif becomes increasingly subtle with increasing sequence length L or decreasing fraction k of the sequences that contain a binding site. This is because the number of competing random motifs increases. Diminishing accuracy of motif detection is thus seen for all tools in Figure 3. However, the accuracy of LocalMotif is consistently higher compared to other tools because LocalMotif's performance depends on localization interval length $|I|$ instead of sequence length L. The localized search reduces competing random motifs, leading to an increase in the accuracy. However, for datasets where motifs are not localized, the comparatively higher accuracy of LocalMotif may not hold.

The accuracy of LocalMotif's interval predictions has been measured in terms of the percentage of overlap between the actual interval, I_a, and predicted interval, I_p. Precisely,

$$\text{overlap percentage} = \frac{|I_a \cap I_p|}{\max(|I_a|, |I_p|)}. \quad (7)$$

The mismatch in the predicted and actual interval lengths is also penalized in this formula by taking the ratio with respect to the longer interval. As seen in Figure 3c, LocalMotif determined the position interval very accurately (overlap $\geq$0.8) in >60% of the cases. This confirms the effectiveness of LocalMotif's scoring function.

3.3 Performance on real datasets

LocalMotif has been further tested to find localized biological motifs in real sequences in three different scenarios: (i) regulatory sequences surrounding the TSS, (ii) segments flanking a known TFBS and (iii) sequences surrounding the peaks in ChIP-Seq data.

3.3.1 Sequences flanking the TSS Promoter sequences surrounding the TSS usually contain highly localized conserved motifs. LocalMotif was tested on insect and mammalian promoter datasets where localized motifs have been previously reported. The insect dataset comprised of 1941 Drosophila core promoter sequences compiled by (Ohler *et al.*, 2002). The sequences are of length 300 bp each aligned −250 to +50 relative to the TSS. Ohler *et al.* determined the core promoter motifs by two separate runs of MEME—one over full 300 bp length, and the other over a sub-interval −60 to +40 relative to the TSS. The full 300 bp length was examined with LocalMotif. As shown in Figure 4, MEME discovered the prominent core promoter motifs only when analyzing the −60 to +40 sub-interval, whereas LocalMotif discovered these motifs given the full 300 bp region. All biologically meaningful motifs reported by LocalMotif had a SCS of 0.14 or above, while two spurious motifs (rows 8 and 9) had SCS <0.06. Thus, SCS allows the discovery of localized core promoter motifs and rejection of spurious motifs. In addition, LocalMotif accurately reported the localization intervals of the motifs which are useful in their identification. For example, the downstream promoter element (DPE) is confirmed as it is found in the (+25, +45) interval.

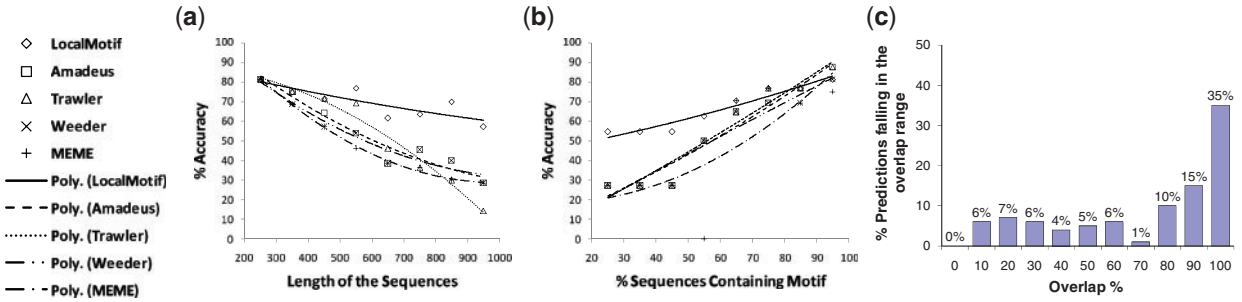

Fig. 3. Performance of LocalMotif, Amadeus, Trawler, Weeder and MEME in simulated short sequence datasets with (**a**) varying sequence length, L, (**b**) varying percentage, k, of sequences containing motif instances. The accuracy of LocalMotif's interval predictions is shown in (**c**).

LocalMotif Results (-250 to +50)						
Rank	*Motif*	*Score*	*RES*	*SCS*	*ORS*	*Position*
1	TCAGTC	1.92	**0.42 (0.0E+00)**	**0.50 (0.0E+00)**	**1.00 (4.7E-08)**	[-5,+15] → Initiator
2	GTCACACT	1.37	**0.43 (0.0E+00)**	**0.23 (0.0E+00)**	**0.71 (3.9E-04)**	[-10,+20] → new motif
3	CTATAAAA	1.27	**0.35 (0.0E+00)**	**0.15 (0.0E+00)**	**0.77 (0.0E+00)**	[-35,-15] → TATA box
4	CAGTTG	1.26	**0.42 (0.0E+00)**	**0.17 (0.0E+00)**	**0.67 (5.1E-06)**	[-5,+15] → Initiator
5	CGGACGTG	1.12	**0.44 (0.0E+00)**	**0.37 (0.0E+00)**	**0.30 (1.2E-01)**	[+25,+45] → DPE
6	CTATCGAT	1.11	**0.40 (0.0E+00)**	**0.14 (0.0E+00)**	**0.57 (1.9E-06)**	[-75, 0] → DRE
7	TCCGTT	0.92	**0.41 (0.0E+00)**	**0.14 (0.0E+00)**	**0.37 (1.2E-05)**	[-5,+15] → Initiator
8	ATATATAT	0.88	**0.32 (0.0E+00)**	**0.02 (1.7E-06)**	**0.54 (1.4E-12)**	[-205,-90]
9	CTCTCTCT	0.86	**0.39 (0.0E+00)**	**0.05 (7.9E-10)**	**0.42 (6.1E-05)**	[-120,-70]
10	GCGTTCGG	0.85	**0.42 (0.0E+00)**	**0.15 (0.0E+00)**	**0.28 (6.8E-04)**	[+10,+40] → DPE

MEME Results (-250 to +50)			MEME Results (-60 to +40)		
Rank	*Motif*	*Score*	*Rank*	*Motif*	*Score*
1	GGTCACACT	5.0e-369 → new motif	1	GGTCACACT	5.1e-415 → new motif
2	CTCTCTC	1.7e-203	2	TATCGATA	1.7e-183 → DRE
3	CGCCGCC	1.1e-151	3	TATAAA	2.1e-138 → TATA box
4	TTTTTTT	1.5e-155	4	TCAGTT	3.4e-117 → Initiator
5	TATCGATA	4.4e-78 → DRE	5	CAGCTG	2.9e-93
6	CAGCCTG	1.5e-80	6	GTATTTT	1.9e-62
7	GGCAACGC	1.4e-55	7	CATCTCT	1.9e-63
8	GTGTGTGT	6.4e-96	8	GGCAACGC	5.1e-29
9	TGCTTTTG	1.2e-39	9	GCGTGCGG	1.9e-12 → DPE
10	GCGCTTTAC	9.5e-24	10	CGAACGGAACG	8.3e-9

Fig. 4. Motifs discovered by MEME and LocalMotif in Drosophila promoters. LocalMotif scores RES, ORS and SCS are shown for each motif with their P-values in the parentheses.

The human dataset included nine different sets of promoters where binding sites for the TFs Oct4, Sox2, Nanog, HNF1A, HNF4A, HNF6, FOXA2, USF1 and CREB1 have been recognized by ChIP-Chip experiments within −8 kb to +2 kb region flanking the TSS (Boyer *et al.*, 2005; Odom *et al.*, 2006). These nine datasets were recently reported to show a sharp peak of the ChIP-Chip signal within 300 bp upstream of the TSS (Koudritsky and Domany, 2008). The full 10 kb region was analyzed for motifs using LocalMotif, Trawler and Amadeus, all of which are capable of handling such large genome-wide datasets. The background model for these tools was constructed from the set of 1 kb upstream promoter and 1 kb exon sequences of all human refseq genes (Supplementary Section C). Each of these tools compares its reported motifs with known PWMs in TRANSFAC. LocalMotif's reported motifs were compared with TRANSFAC using the publically available STAMP tool (Mahony and Benos, 2007) with its default parameters and limiting the similarity E-value to a maximum of 0.001.

In the CREB1 and USF1 datasets, the ChIP TF motif was recognized as the top ranking motif by all tools. In all other datasets, the ChIP TF motif was not recognized as the top ranking motif by any of the tools. This is because in sequences of length 10 kb, the ChIP TF motif is weak compared to random patterns. We studied if the localization of the ChIP TF motifs gives any particular advantage to LocalMotif in being able to rank the ChIP TF motif better than the other tools. In Figure 5, we report the rank of the ChIP TF motif within the results of all the tools tested. If the desired motif was not reported at all, the column is left blank. It is seen that LocalMotif discovered the ChIP TF motif more frequently and with better ranks than the other tools. Most ChIP TF motifs reported by LocalMotif were localized within 1 kb upstream of the TSS (Fig. 5b). LocalMotif also reported a number of other motifs that are clearly localized near the TSS, including Sp1, CAAT box, AP-2, CREB, initiator, E2F,

(**a**)

Dataset	CREB1	FOXA2	HNF1A	HNF4A	HNF6	NANOG	OCT4	SOX2	USF1
# sequences	2108	841	964	4308	1238	1029	372	785	1534
TRAWLER	1	-	-	-	1	-	-	-	1
AMADEUS	1	-		-	-	-	-	-	1
LOCALMOTIF	1	-	7	4	6	8	-	8	1

(**b**)

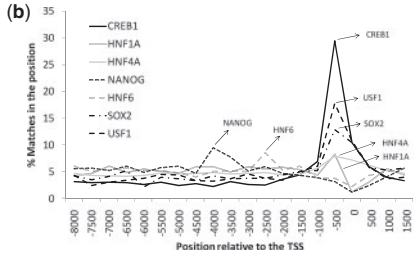

Fig. 5. Results of Trawler, Amadeus and LocalMotif on human promoter datasets: (**a**) table showing whether the ChIP TF motif was reported within the top 10 motifs. (**b**) Distribution around the TSS of motifs reported by Localmotif.

ETS and STAT (Supplementary Section C). Thus, LocalMotif gives advantage in this case by localizing the motif instances in a shorter interval and thus amplifying the motif signal with respect to the random noise.

3.3.2 Sequences flanking a known TFBS An example of co-occurring TFBSs in vertebrate sequences is the close interaction between estrogen receptor (ER) and Forkhead (FoxA1) (Carroll *et al.*, 2005). The dataset in this example consists of 57 ER target sequences from human chromosomes 21 and 22 discovered by ChIP analysis of *in vivo* ER–chromatin complexes (Carroll *et al.*, 2005). Almost all sequences lie distal from the TSS

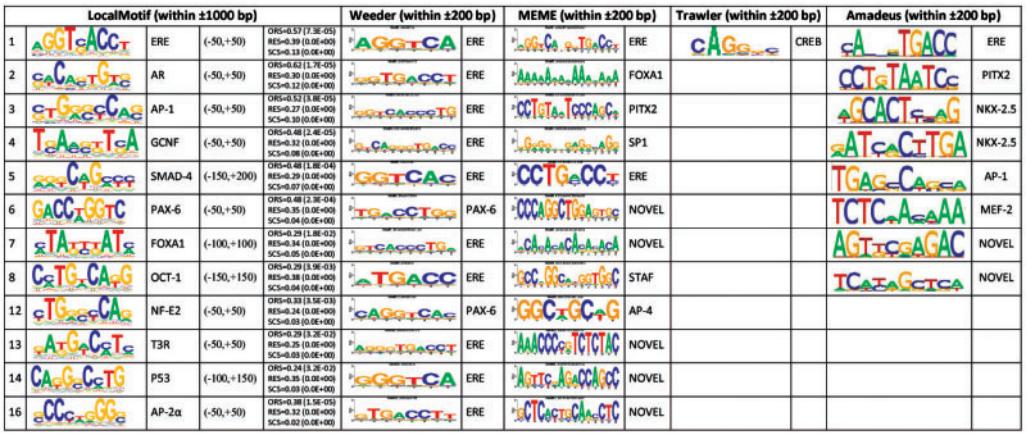

Fig. 6. Motifs discovered by MEME, Weeder, Trawler, Amadeus and LocalMotif in ER ChIP-Seq dataset (Welboren, *et al.*, 2009). LocalMotif scores RES, ORS and SCS are shown for each motif with their *P*-values in the parentheses.

beyond the promoter region and have lengths ranging from 0.2 to 2.5 kb. Thirty-four sequences contain the full ERE motif (length 15 bp, consensus AGGTCANNNTGACCT). The binding sites for Forkhead (consensus TTGTTTNCTT) are experimentally validated proximal to the ER binding sites (Carroll *et al.*, 2005). To verify whether the Forkhead binding adjacent to the ER sites can be discovered *in silico*, the 34 sequences containing full ERE motif were analyzed using MEME, Weeder, Traweler, Amadeus and LocalMotif. The ERE was selected as the anchor point, and its ±500 bp flanking region was analyzed for motifs. The positions of Forkhead binding sites relative to the ERE are shown in Supplementary Section C. Most sites lie close to the ERE. Results of motif finding are reported in the Supplementary Material. Only Amadeus and LocalMotif reported the Forkhead motif with consensus TTTTTTTCTT. About 60% of the experimentally validated Forkhead sites are within the list of sites reported by LocalMotif.

3.3.3 Sequences obtained from ChIP-Seq ChIP-Seq is an emerging high-throughput technology useful for discovering genome-wide *in vivo* binding regions of a TF. The binding regions of the TF are visible as 'peaks' in the ChIP Seq density profile (Johnson *et al.*, 2007). The TFBS usually lies within ±100 bp of the peak maxima. Thus, the ChIP TF motif can be easily discovered in this dataset using any motif finding tool. However, it is more interesting to discover in this dataset the co-regulatory motifs that interact with the ChIP TF. Co-regulatory motifs can occur within 100 bp to 1 kb distance of the peak and show a clear localization around the peak. The LocalMotif scoring function is therefore very useful for discovering co-regulatory motifs in this data. In the present study, two ChIP-Seq datasets were considered.

The first dataset was derived from the recent ChIP-Seq study of 15 TFs in mouse embryonic stem cells (Chen *et al.*, 2008). For each of the 15 ChIP-Seq datasets, the ±200 bp sequences surrounding the 1000 highest intensity peaks were analyzed for motifs. The same background model was used for LocalMotif, Trawler and Amadeus. The background was constructed from 1 kb upstream promoter and 1 kb exon sequences of all mouse refseq genes. The highest ranking motifs reported by Trawler, Amadeus and LocalMotif in each dataset are shown in the Supplementary Section C. Motifs discovered by Amadeus and LocalMotif compared well with the published ChIP TF motifs. Additionally, the LocalMotif motifs with high SCS indicated their concentration around the peak center.

The second dataset comprised of 1000 highest scoring peaks from the ER ChIP-Seq reported in Welboren *et al.* (2009). LocalMotif was used to analyze the ±1 kb region around the peaks. MEME, Weeder, Trawler and Amadeus were used to separately analyze the ±200 bp and ±500 bp regions around the peaks. The background model for LocalMotif, Trawler and Amadeus was constructed from the set of 1 kb upstream promoter and 1 kb exon sequences of all human refseq genes. Weeder was used with its default human background, while MEME did not require a background model. The geometric combination score with equal weights for the three scores was used for LocalMotif. The reported motifs are shown in Figure 6 and their distributions around the center are shown in Supplementary Section C. Results for ±500 bp region are not shown as they are less significant. The reported motifs were compared with TRANSFAC PWMs using the STAMP tool (Mahony and Benos, 2007) with its default parameters and similarity *E*-value cutoff of 0.001. Trawler reported only one motif, the CREB, while Weeder reported only the ERE and Pax-6 motifs. They appear to be less sensitive to weaker motifs when a stronger motif is present. Amadeus and MEME reported more distinct motifs. Amadeus reported ERE, PITX2, NKX-2.5, AP-1, MEF-2 and two novel motifs, while MEME reported ERE, FoxA1, Sp1, STAF, PITX2 and five novel motifs. Only FoxA1, AP-1 and Sp1 have evidence of functioning as co-factors of ER (Carroll *et al.*, 2006; Lin *et al.*, 2007). None of the novel motifs reported by MEME or Amadeus are center-enriched. LocalMotif reported 11 known motifs apart from ERE, all of which except Motif7 (FoxA1) are center enriched around the peaks. Among these, six factors including AP-1, FoxA1, Oct-1, NF-E2, p53 and AR have evidence of functioning as co-factors of ER (Carroll *et al.*, 2006; Lin *et al.*, 2007). Three other motifs, Pax-6, GCNF and T3R are quite similar to ERE. The motifs AP-2α and SMAD-4 appear significant because of the SCS though they have low ORS. Thus using SCS, LocalMotif can discover motifs in ChIP-Seq datasets which are not highly enriched but have localization around the peaks.

4 CONCLUSION

The LocalMotif algorithm has been developed for discovering motifs localized relative to a biological landmark in long regulatory sequences. A localized motif differs from a locally over-represented pattern by the virtue of its spatial confinement within the local interval. A new scoring function called SCS has been developed to measure the spatial confinement. SCS is found to accurately identify the interval of localization. In a sequence set where the motif appears subtle to a usual motif finding algorithm, localization property allows the motif to still be discovered with high accuracy using the LocalMotif scoring function.

Information theoretic framework has been found useful for formulating the scoring function to be consistent for motifs of different lengths and mutations. This allows selection of best motifs while removing their redundant forms. The new formulation gives a clear quantitative as well as qualitative picture of the motif's relevance considering its over-representation, relative entropy and spatial confinement. The three individual scoring measures can be combined in various ways to rank motifs according to their performance in one of more of the three characteristics.

Three specific examples where positional localization of motifs is found useful were reported. These include regulatory sequences surrounding the TSS, sequences flanking a known TFBS and sequences flanking the peaks in a ChIP-Seq dataset. LocalMotif could detect motifs in longer sequences as compared to other tools which are sensitive to sequence length, and could amplify weak motifs in case they were localized. For similar reasons, LocalMotif also detected a number of co-regulatory motifs flanking a main motif. With the emergence of ChIP-Seq, co-motif discovery using positional localization is extremely relevant. The interval predictions reported by LocalMotif provide additional insight into the range of TF–TF interactions.

LocalMotif is presently based on the (l, d) motif model. There are emerging opinions in the literature to give a more accurate description of the motif, such as gapped motifs, or motifs based on IUPAC character set. It will be interesting to study which motif representations lead to more accurate results on biological data. The LocalMotif algorithm could also be improved from a heuristic search to exact search in the future using efficient data structures such as suffix tree and FM index.

Funding: This work is supported by Academic Research Fund (AcRF) R-252-000-326-112.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Boyer,L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Buhler,J. and Tompa,M. (2002) Finding motifs using random projections. *J. Comput. Biol.*, **9**, 225–242.

Carroll,J.S. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

Carroll,J.S. *et al.* (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.*, **38**, 1289–1297.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Eskin,E. and Pevzner,P.A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, **18** (Suppl. 1), S354–S363.

Ettwiller,L. *et al.* (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.

Fratkin,E. *et al.* (2006) MotifCut: regulatory motifs finding with maximum density subgraphs. *Bioinformatics*, **22**, e150–e157.

Friberg,M. *et al.* (2005) Scoring functions for transcription factor binding site prediction. *BMC Bioinformatics*, **6**, 84.

Henikoff,S. *et al.* (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **163**, GC17–GC26.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Keich,U. and Pevzner,P.A. (2002a) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374–1381.

Keich,U. and Pevzner,P.A. (2002b) Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics*, **18**, 1382–1390.

Koudritsky,M. and Domany,E. (2008) Positional distribution of human transcription factor binding sites. *Nucleic Acids Res.*, **36**, 6795–6805.

Lawrence,C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lin,C.Y. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet.*, **3**, e87.

Linhart,C. *et al.* (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.

Liu,X. *et al.* (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.

Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.

Marsan,L. and Sagot,M.F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.*, **7**, 345–362.

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

Molina,C. and Grotewold,E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.

Odom,D.T. *et al.* (2006) Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.*, **2**, 2006 0017.

Ohler,U. *et al.* (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, 0087.

Pavesi,G. *et al.* (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.

Pevzner,P.A. and Sze,S.H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 269–278.

Qi,Y. *et al.* (2006) High-resolution computational models of genome binding events. *Nat. Biotechnol.*, **24**, 963–970.

Roth,F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Sinha,S. and Tompa,M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.

Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

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

Tharakaraman,K. *et al.* (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics*, **21** (Suppl. 1), i440–i448.

Thijs,G. *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.

Vardhanabhuti,S. *et al.* (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.

Welboren,W.J. *et al.* (2009) ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.*, **28**, 1418–1428.

Wierstra,D. *et al.* (2008) Fitness expectation maximization. In *Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN X*. Springer, Dortmund, Germany, pp. 337–346.