BIOINFORMATICS

Vol. 15 nos 7/8 1999
Pages 622–630

Estimating transcription factor bindability on DNA

Tatsuhiko Tsunoda and Toshihisa Takagi

Genome Data Base, Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Received on October 30, 1998; revised on April 1, 1999; accepted on April 13, 1999

Abstract

Motivation: *Precise analysis of the genetic network, gene function and transcription regulation requires accurate prediction of transcription factor (TF) bindability on DNA. For calculating the matching score between an input sequence and a set of known TF binding sites, we use positional weight matrices (PWMs) and Bucher's calculating method (Bucher,* J. Mol. Biol., ***212****, 563–578, 1990). Since estimating TF binding sites requires cut-off values, we propose a robust cut-off value determining algorithm.*

Results: *We generalize the concept of local over-representation with statistics, and propose a new algorithm for determining the cut-off value using the background rate estimated on non-promoters. The algorithm iteratively determines parameters separating instances into phenomena-dependent and phenomena-independent subsets. Our system includes the method of re-estimating cut-off values of TFs that mis-recognize other TF preferred regions. Our data source comprised 433 non-redundant vertebrate promoters including viral promoters, from Eukaryotic Promoter Database (EPD) R.50. The method is applied to 205 vertebrate TFs that have frequency matrices in TRANSFAC Ver.3.4 and the cut-off values of all of them can be determined.*

Availability: *The cut-off values and TF binding site predicting tool are available at http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/TFBIND. We also provide the cut-off value estimating programs.*

Contact: *tatsu@ims.u-tokyo.ac.jp*

Introduction

Precise prediction of TF (transcription factor)–DNA interaction is an effective clue for estimating interaction between TFs (Tsunoda and Takagi, 1998), predicting promoter regions on DNA (Fickett and Hatzigeorgiou, 1997), estimating gene function by analyzing organization specificity of its upstream regulatory region (Fickett and Hatzigeorgiou, 1997; Fujibuchi and Kanehisa, 1997) and reconstructing genetic networks.

To identify TF binding sites on DNA, we have to specify each DNA-binding motif [e.g. positional weight matrices (PWMs)], calculate matching scores between the motif and input sequences, and separate them from non-binding sites with a cut-off value (Frech *et al*., 1997). Several researchers provide methods for defining the PWMs (Stormo and Hartzell, 1989; Goodrich *et al*., 1990; Hertz *et al*., 1990; Laurence *et al*., 1993; Hertz and Stormo, 1995; Neuwald *et al*., 1995; Quandt *et al*., 1995; Wolfertstetter *et al*., 1996), and we can currently obtain the matrix databases MD (Chen *et al*., 1995) and TRANSFAC (Heinemeyer *et al*., 1998). To search for TF binding signals, several algorithms have been proposed: MATRIX SEARCH (Chen *et al*., 1995), ConsInspector (Frech *et al*., 1993), MatInspector (Quandt *et al*., 1995), TFSearch, TESS search, etc. However, since they provide cut-off values without clear measure, they sometimes produce many false positives or false negatives.

Bucher (1990) proposed a novel algorithm for detecting the cut-off value of the binding score for extracting its motif, and for identifying its preferred binding region in promoters. However, it has limitations (see Discussion). To avoid these, we modify the concept of local window, generalize the concept of local overrepresentation, use non-promoter sequences to estimate the background instead of promoters, and propose an algorithm for determining the cut-off value for the given PWM of each TF.

Algorithm

New concepts

Let us suppose we have a set $\boldsymbol{S} = \{P_1, P_2, \ldots, P_{N_0}\}$, where each $P_k (1 \leq k \leq N_0)$ defines an independent promoter sequence. We align all promoters in $\boldsymbol{S}$ with transcription start sites (TSS), and define x as the position relative to the TSS on each promoter (Figure 1).

We introduce a local window of width w (bp) that centers x bp upstream of the TSS (Figure 1). Let us suppose we assign a different cut-off value th for the binding score of the PWM for each TF (see Appendix A) respectively. We define $N(S, f, th, x, w)$ as the number of promoters in $\boldsymbol{S}$ that bind TF f using cut-off value th inside this window (see Figure 1 for example).

Supposing N non-promoter sequences are aligned randomly, we estimate the background rate $BG(f, th, N, w)$ as the number of the sequences that bind TF f within an arbitrary window of size w using cut-off value th. We also define $SD(f, th, N, w)$ as the standard deviation calculated with the background rate (see Appendix B).

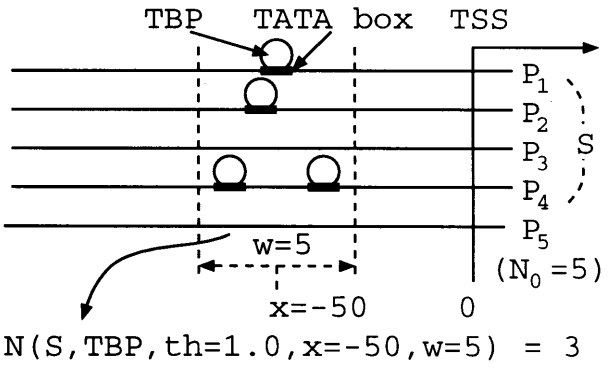

Fig. 1. Promoter alignment, local window and example of TF binding sites.

We introduce generalized local overrepresentation:

$$O_g\ (\boldsymbol{S}, f, th, x, w) = \frac{N(\boldsymbol{S}, f, th, x, w) - BG(f, th, N0, w)}{SD(f, th, N0, w)}$$

This shows the significance of the detected number of promoters that bind the TF compared with the random fluctuation of all instances. When $O_g > O_c$ (the cut-off value for detecting local overrepresentation, e.g. 2.0) and $N(\boldsymbol{S}, f, th, x, w) > N_c$ (the minimum number of coincidences for detecting local overrepresentation, e.g. 4), we consider that there is significant local overrepresentation.

Here we consider one TF. In general, we define the optimum cut-off value that can correctly discriminate functional sites from background sequences. However, functional sites are not given explicitly. Bucher's method extracts TF motifs and determines optimum cut-offs by maximizing the ratio of signal to noise at a preferred region according to the assumption that such functional sites are conserved at the region (Bucher, 1990). The preferred region is detected by estimating the local overrepresentation. We also use the information on local overrepresentation to determine cut-off values.

However, if we use the sequences within the preferred region in all promoters considering they are functional sites, there will be a problem that the estimated cut-off value will be lower than the actual value since the sequences still include both functional sites and non-functional sites; the full set of promoters ($\equiv \boldsymbol{S}$) consists of two types of promoters: promoters in which the TF functional sites are conserved in the preferred region during evolution ($\equiv \boldsymbol{S}_a$), and other promoters in which such functional sites are not conserved ($\equiv \boldsymbol{S}_b$) since they were not required. To avoid the above problem, we must discriminate $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$ correctly, which is also a problem that we must solve. Since we do not know any of the characteristics for the promoter in $\boldsymbol{S}$ initially, no explicit information for discriminating $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$ is given here. To discriminate $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$, we must determine the cut-off beforehand. That is, the processes of determining the cut-off and discriminating $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$ are mutually dependent. Thus, we start from the situation that the cut-off value ($\equiv th$) has been determined as some value. We will change this value later.

Using *th*, we can calculate TF binding sites in promoters. If we find many promoters that have TF binding sites within a window, we can consider they will be functional sites of the TF. We separate $\boldsymbol{S}$ into two subsets: $\boldsymbol{S}_a$, in which promoters have TF binding sites within the window, e.g. TATA-containing promoters, and $\boldsymbol{S}_b$, which do not. $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ are fixed. Here, we check whether these sets satisfy the following condition: $\boldsymbol{S}_b$ does not have local overrepresentation within the window even if TF binding sites are re-estimated at any hypothetical threshold *th′* lower than *th*. In $\boldsymbol{S}_b$, if by temporarily lowering the cut-off value, statistically significant local overrepresentation is detected, then there is evidence that it consists of functional sites. However, this is opposed to the definition of $\boldsymbol{S}_b$; it means that some promoters that should be classified into $\boldsymbol{S}_a$ are misclassified into $\boldsymbol{S}_b$. We can conclude that the cut-off *th′*, which was used to discriminate $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$, is too high. Thus, we must reduce *th′* with some step, separate $\boldsymbol{S}$ into $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$, and recheck. We repeat this process until $\boldsymbol{S}_b$ does not have local overrepresentation within the window even if TF binding sites are re-estimated at any hypothetical threshold *th′* lower than *th*.

Although we considered one window above, preferred regions can be found multiply and anywhere in the promoters. Thus, we consider the cut-off to be optimum if it satisfies the two following conditions:

1. Anywhere in the promoters, $\boldsymbol{S}_b$ does not have local overrepresentation within the window even if TF binding sites are re-estimated at any hypothetical threshold *th′* lower than *th*.
2. Maximum cut-off that satisfies 1.

If (1) is not satisfied, we can consider that some promoters that should be classified into $\boldsymbol{S}_a$ are misclassified into $\boldsymbol{S}_b$. We cannot use the cut-off value since it discriminates $\boldsymbol{S}_a$ from $\boldsymbol{S}_b$ wrongly. If (2) is not satisfied, it means that the actual cut-off value is lower than that determined above. We can conclude that, although there are binding sites whose scores are between the actual cut-off value and that determined, we cannot detect them as local overrepresentation. However, this is opposed to an assumption that the distribution of the binding scores of functional sites starts from the actual cut-off value with a statistically significant level. The basis for this assumption is that the binding score might be becoming lower by mutation, etc., while they have binding ability. By these considerations, we consider the cut-off value that satisfies the above conditions as optimum.

Table 1. Example of background rate and standard deviation [for TBP (TATA box) and w = 5]

		BG(*TBP*, *th*, *N*, 5)					*SD*(*TBP*, *th*, *N*, 5)				
th	*N*	1	2	3	4	5	1	2	3	4	5
0.8		0.10	0.20	0.30	0.40	0.50	0.30	0.45	0.55	0.63	0.71
1.0		0.06	0.13	0.19	0.25	0.31	0.24	0.36	0.44	0.50	0.56

```
1   Iterative Algorithm
2   INPUT: S, PWM_f
3   OUTPUT: optimum cut-off value for f
4   begin
5     align S with TSS; th := 1.0;
6     for x := x_min to x_max do
7       for w := w_min to w_max do
8         begin
9           repeat
10            Search signals of f on P_k in S using
11            PWM_f and th within the window;
12            Separate S into S_a and S_b;
13            th' := th;
14            repeat
15              th' := th' − step;
16              Search signals of f on P_k in S_b using
17              PWM_f and th' within the window;
18              Count N(S_b, f, th', x, w);
19              Calculate O_g(S_b, f, th', x, w);
20              if O_g > O_c and N > N_c then LOR is
21              detected in S_b;
22            until th' < 0 or LOR is detected in S_b
23            if LOR is detected in S_b
24              then th = th − step;
25          until LOR is not detected in S_b
26        end ;
27  end ;
```

Fig. 2. Iterative algorithm for determining cut-off values. LOR means local overrepresentation.

Iterative algorithm for estimating cut-off values

The following procedures are applied to every factor respectively. The simplified code is shown in Figure 2. Here, we focus on one TF *f*, e.g. TBP.

1. Initialization: Suppose O_c = 2.0, N_c = 1 (fixed for any situation), and *BG* and *SD* are as shown in Table 1. Align all promoters with the TSS, and set the cut-off value *th* to 1.0 (line 5).
2. Setting the window: lines 6, 7. Consider window W of width w at position x, e.g. w = 5 and x = –50 (Figure 1).
3. Searching for signals: lines 10, 11. Search all candidates of the binding sites of *f* on all promoters within the window using *th* (Figure 1).
4. Separating the promoter set: line 12. For example, let the set of promoters TBP be bindable within the window as $\boldsymbol{S}_a$ = {P_1, P_2, P_4}, and the others as $\boldsymbol{S}_b$ = {P_3, P_5} (Figure 1).
5. Next, we check whether the set $\boldsymbol{S}_b$ has local overrepresentation potential by the following steps:

(a) Search all candidates of the binding sites of *f* on promoters in set $\boldsymbol{S}_b$ using cut-off value *th*′ (temporary value for checking) within the window (lines 16, 17). For example, for the two promoters in $\boldsymbol{S}_b$, the TBP binding sites are estimated with *th*′ (e.g. 0.8), which is lower than *th*. According to Table 1, the background rate is 0.2 for two promoters ($\boldsymbol{S}_b$), and the standard deviation 0.45. Within the window, the number of promoters (in $\boldsymbol{S}_b$) that bind *f* is counted (line 18).

(b) If, for example, the number of *f*-binding promoters is 2, then $O_g = \frac{2 - 0.2}{0.45} = 4.0$ (line 19). Because it satisfies the conditions that $O_g > O_c$ and $N > N_c$, the set $\boldsymbol{S}_b$ has local overrepresentation potential; this set does not satisfy the criterion that $\boldsymbol{S}_b$ does not have local overrepresentation within the window even if TF binding sites are re-estimated at any hypothetical threshold *th*′ lower than *th*. Reduce *th* with some step size, e.g. 0.01, and go to (3) (line 23, 24).

(c) If the result of 5(a) is 1, then $O_g = \frac{1 - 0.2}{0.45} = 1.8$. Since $O_g < O_c$ [as the background rate is the same as 5(b)], we do not detect any local overrepresentation potential with *th*′ within this window.

6. If $\boldsymbol{S}_b$ does not have local overrepresentation within the window even if TF binding sites are re-estimated at any hypothetical threshold *th*′ lower than *th*, go to the next step (line 25).

Thus, the final cut-off value is the minimum one calculated by this procedure for all windows ($w_{min} \le w \le w_{max}$, x in entire region).

Decision of preferred region using statistical significance

With the cut-off values determined using the preceding algorithm, choose w ($w_{min} \le w \le w_{max}$) that maximizes $O_g(\boldsymbol{S}, f, th, x, w)$ at each x and define $O_g^0(x)$ as follows (from here, we omit *f*, $\boldsymbol{S}$ and *th*):

$$O_g^0(x) = \max_{w_{\min} \le w \le w_{\max}} O_g(\boldsymbol{S}, f, th, x, w).$$

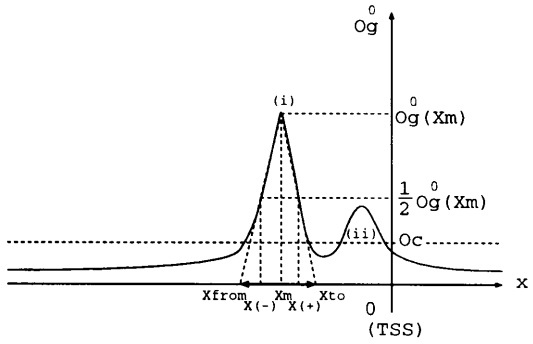

Fig. 3. Decision of preferred region using statistical significance.

Suppose the plot is like that in Figure 3.

$$O_g^0(x_m) = \max_{x \text{ in region}} O_g^0(x)$$

if $O_g^0(x_m) \leq O_c$ **then** end **else** goto next step.

$$x_{(+)} - x_m = \min_{O_g^0(x_i) = \frac{1}{2}O_g^0(x_m), x_i > x_m} x_i - x_m$$

$$x_m - x_{(-)} = \min_{O_g^0(x_i) = \frac{1}{2}O_g^0(x_m), x_i > x_m} x_m - x_i$$

We define the region between $x_{from}(= 2x_{(-)} - x_m)$ and $x_{to}(= 2x_{(+)} - x_m)$ as one of the preferred regions of this factor. To locate other preferred regions [e.g. (ii) in Figure 3], delete the preferred region [e.g. (i)] and repeat the procedure.

Re-estimation of the cut-off value of TF that mis-recognizes other TF binding sites

We can sometimes find several TFs whose consensus sequences are similar. They may bind with overlapping sites or exclusively to the same sites. However, not all of them will be functional. For example, transcription factor GATA is not functional when binding to every TATA box. For determining cut-off values, it may be harmful to use such mis-recognized sites for our algorithm since it may misclassify the promoters into $\boldsymbol{S}_a$, and the cut-off value will be lower than the real value. The most extreme case is that, inside one window, our algorithm mis-recognizes a local overrepresentation of another TF to be that of the target TF, although its binding sites are not conserved. To avoid this, we want to determine cut-off values using sites above suspicion. Thus, we propose an additional procedure for identifying such mis-recognition of preferred regions of other TFs and masking such sites.

We define $PR^i(x^i_{from}, x^i_{to})$ as an extracted preferred region of F^i. We assume that F^B mis-recognizes F^A's preferred region as F^B's binding sites when all the following conditions are satisfied:

1. Overlap of preferred region: $x^A_{from} - \Delta x^A \leq x^B_{from} \leq x^A_{from} + \Delta x^A$ and $x^A_{to} - \Delta x^A \leq x^B_{to} \leq x^A_{to} + \Delta x^A$, where $\Delta x^A = \max(a(x^A_{to} - x^A_{from}), b)$, a is the critical overlapping ratio and b is the maximum shift of position; both are given.
2. Overlap of promoter sets: $n(\boldsymbol{S}^{AB}) \geq c \cdot n(\boldsymbol{S}^A)$ or $n(\boldsymbol{S}^{AB}) \geq c \cdot n(\boldsymbol{S}^B)$, where $\boldsymbol{S}^A$ is defined as the set of promoters that bind F^A within PR^A, the number of which is $n(\boldsymbol{S}^A)$. $\boldsymbol{S}^B$ is defined as the set of promoters that bind F^B within PR^A, the number of which is $n(\boldsymbol{S}^B)$. $\boldsymbol{S}^{AB} \equiv \boldsymbol{S}^A \cap \boldsymbol{S}^B$. c $(0 < c \leq 1)$ is assumed to be given.
3. The number of common promoters $[n(S^{AB})]$ is more than a lower limit n_c, which is given.

Secondly, to avoid mis-recognition of F^B as its binding site on each promoter, mask F^A's binding sites within PR^A whose scores are above the estimated cut-off value for F^A.

Finally, apply the cut-off value determining algorithm again for the positions which are not masked.

Implementation

We used 205 vertebrate TFs from the database TRANSFAC Ver.3.4 (Heinemeyer *et al.*, 1998). Using the PWM data, each TF matching score (binding score) in each position was calculated according to Bucher's method (Bucher, 1990) (see Appendix A). The smoothing parameter for it is 0.01. We used EPD R.50 (Bucher and Trifonov, 1986) for promoter sequences. Non-redundant sequences were taken from the vertebrate promoters (including viral promoters). Our final set consisted of 433 promoters for which the region –349 to +100 bp of the TSS had been determined. From GenBank, we extracted sequences totalling 664 505 bp (1 329 010 bases) according to the list of non-promoters from Dr Prestridge at Minnesota University (Prestridge, 1995).

We set the following parameters: $w_{min} = 1$, $w_{max} = 9$, $O_c = 2.0$, $N_c = 4$, and the step size for reducing cut-off value = 0.01. We also set the parameters for detecting TFs that mis-recognize the preferred regions of other TFs: $a = 0.1$, $b = 10$, $c = 0.75$, $n_c = 30$.

The TATA box and cap sites are estimated only on the sense strand since they are known to be orientation specific. Each binding site score for the other TFs is calculated for both strands.

The program is written in C++ and is executable on a supercomputer (Sun Enterprise 10000, 64 processors) at the Human Genome Center.

Results

The estimated cut-off value, background rate and recall of the experimental data using the cut-off value of each TF are shown in Table 2. For indicating the factor, we used the notation of TRANSFAC matrix entry (Heinemeyer *et al.*, 1998). First, they have an identifier that indicates vertebrates (V$), followed by an acronym for the factor, and a consecutive number discriminating between different matrices for the same factor. Secondly, instead of the consecutive number, those TF matrices which have been generated from TRANSFAC site entries connected to a certain transcription factor, IDs end up with an abbreviation of the least quality of the sites used to construct the matrix. Finally, a matrix with an ID like V$AP_C has been derived from a 'consensus description' constructed with the aid of ConsIndex (Frech *et al.*, 1993). The entry indicating GC box (M00255; V$GC_01) was originally collected data by Bucher using his computational method (Bucher, 1990). While the frequency matrices of Sp1 (M00008; V$SP1_01 and M00196; V$SP1_Q6) were compiled from experimentally determined data, TRANSFAC includes these matrices as separated entry, each cut-off of which we determined respectively. The fourth column in Table 2 shows the preferred region that includes the maximum O_g^0 in the spectrum of each TF binding site. We call them the most preferred regions. The fifth column presents the number of promoters that have TF binding sites within the most preferred region. The cut-off values of the TATA box, GC box and CCAAT box are lower than Bucher's, while the cut-off value of the cap site is higher than his. The estimated background rates are shown in the sixth column in Table 2.

We found 36 TFs that mis-recognize the TATA box (indicated by 'o' in front of the ACCESS number in Table 2) following the general rule that these sites are indeed the TATA box. We re-estimated the cut-off values of these TFs using both TATA-containing promoters on which each TATA box is masked and TATA-less promoters unchanged. We are still concerned that some TFs mis-recognize other TF sites. For example, transcription factor AP2 may mis-recognize the preferred region of transcription factor Sp1. However, since the estimated preferred region of AP2 is wider than the preferred region of Sp1, we considered them independent.

To check the appropriateness of the estimated cut-off values, we applied them to the sequences compiled in TRANSFAC (Heinemeyer *et al.*, 1998). The factor which binds to each sequence element is given with its 'quality' value, ranging from 1 to 6, and reflecting the experimental reliability of a certain protein–DNA interaction [1, functionally confirmed factor binding sites; 2, binding of pure protein (purified or recombinant); 3, immunologically characterized binding activity of a cellular extract; 4, binding activity characterized via a known binding sequence; 5, binding of uncharacterized extract protein to a bone fide element; 6, no quality assigned]. From columns 7 to 12 in Table 2, Q1 includes only data of quality 1, and Q1-n includes data from quality 1 to quality *n*. Because the data set is small and it is uncertain whether each site is actually functional, we used the data as a rough check. Fifty TFs completely lack data. From Table 2, we can conclude the recall is almost acceptable. In several cases where the recall is rather low (50% or less), we can suggest several reasons: too little data and limitation of our algorithm (see Discussion).

Discussion

The Bucher method (Bucher, 1990) also aligns many promoters with the TSS and uses information on local overrepresentation. With clues as to the motif (minimum string) and initial values of parameters (lower bound and upper bound of the width, etc.) entered in manually, it iteratively calculates the motif, preference region and cut-off value for the PWM of each TF. The basic strategy is to maximize the difference between the signal and background ratios. Because it automatically discriminates the signal from background by a maximizing procedure on promoters, an explicit background rate is not required. The result is correct when the sequences are clearly separated into signal and other areas, and when the separation is true. Although we can find TFs, e.g. Sp1, having several preference regions widely spread over the promoter, only one of them is discriminated as the signal area from the other regions, which are recognized as background. Hence, the amount of background estimated is too high, which leads to excessive estimation of its binding cut-off value, and the preference region is too narrow.

Our algorithm relies on TF frequency matrices in TRANSFAC for calculating PWMs. It is not targeted to optimize them. In our method, because the background rate is estimated on the non-promoters, and because the criteria of local overrepresentation have been set at each position over the full area, it avoids the above mistake. In fact, the preferential site is not clear, and not specified as singular; therefore, we took an approach that the local overrepresentative promoters are determined locally. Our system checked local windows, including every window size at each position, because it can vary according to the binding density by the width and position of the window. Our technique is original since it provides generality and the ability to avoid mis-recognition. Our algorithm was applied to 205 vertebrate TFs that have frequency matrices in TRANSFAC and the cut-off value of all of them can be determined. This can be done without any manual tuning of initial parameters. Our criteria can be applied to other types of problems, e.g.

Table 2. Final results of the estimated cut-off value, background rate (/1*bp* · 1*sequence*), and recall of the experimental data using the cut-off value for each TF

	ACCESS	FACTOR	CUT-OFF	Pref. reg.	#pro	background	Q1	Q1-2	Q1-3	Q1-4	Q1-5	Q1-6
	M00189	V$AP2_Q6	0.78	-173 - -36	391	0.0269						
	M00008	V$SP1_01	0.78	-69 - -35	323	0.0297	100.0	99.1	99.2	99.3	99.5	99.5
	M00252	V$TATA_01	0.77	-40 - -23	297	0.0065		100.0	66.7	66.7	66.7	91.3
	M00255	V$GC_01	0.78	-74 - -45	292	0.0243						
	M00196	V$SP1_Q6	0.75	-78 - -37	273	0.0144	93.8	97.4	97.2	96.6	96.0	96.4
	M00216	V$TATA_C	0.74	-40 - -24	261	0.0077		100.0	66.7	66.7	71.4	75.6
	M00175	V$AP4_Q5	0.78	32 - 65	250	0.0175						
	M00084	V$MZF1_02	0.81	-99 - -24	248	0.0097		100.0	100.0	100.0	100.0	100.0
	M00176	V$AP4_Q6	0.76	-31 - -3	208	0.0147						
	M00085	V$ZID_01	0.76	7 - 68	206	0.0071						
	M00244	V$NGFIC_01	0.72	-80 - -38	204	0.0088						100.0
	M00185	V$NFY_Q6	0.77	-98 - -59	181	0.0087						
	M00253	V$CAP_01	0.87	-5 - 6	179	0.0226						
	M00254	V$CAAT_01	0.78	-105 - -70	174	0.0093						
	M00083	V$MZF1_01	0.83	-66 - -27	165	0.0089		100.0	100.0	100.0	100.0	100.0
	M00243	V$EGR1_01	0.74	-83 - -36	152	0.0049		90.0	90.9	91.7	92.9	95.7
	M00115	V$TAXCREB_02	0.61	16 - 33	139	0.0165	90.0	92.5	93.6	94.5	94.5	93.8
	M00114	V$TAXCREB_01	0.71	-129 - -84	139	0.0043	90.0	97.5	91.5	92.7	92.7	95.1
	M00227	V$VMYB_02	0.78	-7 - 18	138	0.0092		100.0	100.0	100.0	100.0	100.0
	M00246	V$EGR2_01	0.74	-81 - -37	137	0.0050	100.0	100.0	100.0	100.0	100.0	100.0
	M00072	V$CP2_01	0.78	61 - 78	127	0.0145		100.0	100.0	100.0	100.0	88.9
	M00050	V$E2F_02	0.74	-259 - -236	116	0.0088			100.0	100.0	100.0	100.0
	M00180	V$E2F_Q6	0.73	54 - 87	113	0.0046			100.0	100.0	100.0	100.0
	M00051	V$NFKAPPAB50_01	0.75	-108 - -71	109	0.0049		100.0	100.0	100.0	100.0	100.0
	M00245	V$EGR3_01	0.74	-77 - -36	108	0.0035						
	M00001	V$MYOD_01	0.79	-70 - -48	105	0.0094	100.0	100.0	100.0	100.0	100.0	94.7
	M00005	V$AP4_01	0.75	-32 - -8	93	0.0069		100.0	100.0	100.0	100.0	100.0
	M00143	V$PAX5_01	0.76	22 - 40	88	0.0072				100.0	100.0	100.0
	M00113	V$CREB_02	0.77	-230 - -210	87	0.0076	100.0	100.0	97.6	98.0	98.0	98.6
	M00172	V$AP1FJ_Q2	0.81	-103 - -73	85	0.0053						
	M00004	V$CMYB_01	0.74	-232 - -206	83	0.0049	50.0	57.1	57.1	57.1	57.1	66.7
	M00237	V$AHRARNT_02	0.71	30 - 58	82	0.0043	100.0	100.0	100.0	100.0	100.0	100.0
	M00273	V$R_01	0.72	-16 - 19	81	0.0029		100.0	100.0	100.0	100.0	100.0
	M00139	V$AHR_01	0.71	-63 - -40	79	0.0054	100.0	100.0	100.0	100.0	100.0	100.0
o	M00003	V$VMYB_01	0.77	-326 - -300	79	0.0103		100.0	100.0	100.0	100.0	100.0
	M00209	V$NFY_C	0.76	-92 - -57	78	0.0022		50.0	50.0	57.1	50.0	85.3
	M00235	V$AHRARNT_01	0.76	14 - 32	76	0.0066	100.0	100.0	100.0	100.0	100.0	93.3
	M00141	V$LYF1_01	0.82	-82 - -71	76	0.0137	100.0	100.0	100.0	100.0	100.0	100.0
	M00002	V$E47_01	0.77	18 - 33	73	0.0080		100.0	100.0	100.0	100.0	100.0
	M00122	V$USF_02	0.75	-28 - -17	69	0.0111	100.0	87.5	91.3	94.9	95.2	95.5
o	M00127	V$GATA1_03	0.78	-11 - -1	64	0.0208	96.4	95.0	95.2	97.3	97.7	97.7
	M00058	V$HEN1_02	0.71	-4 - 32	64	0.0026		100.0	100.0	100.0	100.0	100.0
	M00184	V$MYOD_Q6	0.77	-4 - 4	63	0.0138						
	M00147	V$HSF2_01	0.79	-94 - -83	62	0.0092	100.0	100.0	100.0	100.0	100.0	100.0
	M00264	V$STAF_02	0.75	-74 - -55	60	0.0051						
	M00271	V$AML1_01	0.83	77 - 89	59	0.0093		100.0	100.0	100.0	100.0	100.0
	M00057	V$COMP1_01	0.77	-19 - -7	58	0.0084						
	M00032	V$CETS1P54_01	0.81	-272 - -261	58	0.0093		100.0	100.0	100.0	100.0	100.0
o	M00278	V$LMO2COM_02	0.79	-13 - 1	56	0.0172			100.0	100.0	100.0	100.0
	M00277	V$LMO2COM_01	0.78	-79 - -72	56	0.0123			100.0	100.0	100.0	100.0
	M00098	V$PAX2_01	0.75	-102 - -86	56	0.0066						0.0
	M00055	V$NMYC_01	0.74	-309 - -298	55	0.0076						
	M00257	V$RREB1_01	0.78	-143 - -117	54	0.0055						
	M00056	V$MYOGNF1_01	0.74	-111 - -93	54	0.0047		100.0	100.0	100.0	100.0	100.0
	M00025	V$ELK1_02	0.78	-280 - -265	53	0.0055		100.0	100.0	100.0	100.0	100.0
	M00108	V$NRF2_01	0.77	14 - 24	52	0.0076						100.0
	M00280	V$RFX1_01	0.75	38 - 53	51	0.0054						100.0
	M00017	V$ATF_01	0.74	-17 - -3	51	0.0056						
	M00177	V$CREB_Q2	0.75	-14 - -4	50	0.0067						
	M00262	V$STAF_01	0.72	-76 - -61	48	0.0046						
	M00121	V$USF_01	0.74	-70 - -46	48	0.0030	100.0	81.2	78.3	79.5	78.6	79.5
	M00178	V$CREB_Q4	0.74	-54 - -45	46	0.0075						
	M00158	V$COUP_01	0.80	-217 - -187	45	0.0027	100.0	100.0	100.0	100.0	100.0	100.0
	M00155	V$ARP1_01	0.75	-29 - -19	44	0.0071	0.0	87.5	87.5	87.5	87.5	91.7
	M00187	V$USF_Q6	0.79	-5 - 1	42	0.0097						
	M00075	V$GATA1_01	0.77	-5 - -3	42	0.0249	84.6	88.2	89.5	90.0	92.1	92.1
	M00181	V$E2_Q6	0.74	6 - 22	41	0.0039						
	M00099	V$S8_01	0.76	-236 - -222	41	0.0071						
	M00053	V$CREL_01	0.81	72 - 84	41	0.0057		100.0	100.0	100.0	100.0	100.0

protein–DNA interaction, protein–protein interaction, ligand/domain docking, etc.

There may be cases where our algorithm cannot deal with factors whose binding sites are sparsely spread over the promoters. For instance, when the coincidence rate is within the range of the random coincidence level locally, it might exceed the random level globally. Such a positional shift of binding sites may arise from flexible distance between interacting TFs and more complex mechanisms such as dynamic structural changes (e.g. DNA bending). Since we set

Table 2. *Continued*

	ACCESS	FACTOR	CUT-OFF	Pref. reg.	#pro	background	Q1	Q1-2	Q1-3	Q1-4	Q1-5	Q1-6
	M00071	V$E47_02	0.76	18 - 29	39	0.0048		100.0	100.0	100.0	100.0	100.0
	M00239	V$T3R_01	0.73	-15 - -7	38	0.0072	100.0	80.0	80.0	80.0	80.0	87.5
	M00193	V$NF1_Q6	0.79	-128 - -112	38	0.0032						
	M00281	V$RFX1_02	0.75	-314 - -298	37	0.0036						100.0
	M00160	V$SRY_02	0.76	-312 - -309	37	0.0190		100.0	100.0	100.0	100.0	100.0
	M00076	V$GATA2_01	0.78	-5 - -3	37	0.0234						
	M00052	V$NFKAPPAB65_01	0.76	-110 - -96	37	0.0046	100.0	100.0	100.0	100.0	100.0	100.0
	M00118	V$MYCMAX_01	0.71	-200 - -184	36	0.0030	100.0	100.0	100.0	100.0	100.0	91.7
	M00272	V$P53_02	0.79	-164 - -160	35	0.0134			100.0	100.0	100.0	75.0
	M00269	V$XFD3_01	0.78	-42 - -32	35	0.0055						
o	M00126	V$GATA1_02	0.77	-12 - 2	34	0.0115	92.9	95.0	95.2	93.2	93.0	93.0
	M00074	V$CETS1P54_02	0.83	-278 - -271	34	0.0063						
	M00066	V$TAL1ALPHAE47_01	0.76	-309 - -295	34	0.0040		100.0	100.0	100.0	100.0	100.0
	M00208	V$NFKB_C	0.75	73 - 86	33	0.0038	100.0	100.0	100.0	100.0	100.0	100.0
	M00080	V$EVI1_03	0.71	-40 - -31	33	0.0055						100.0
	M00077	V$GATA3_01	0.82	-289 - -284	33	0.0096	100.0	100.0	100.0	100.0	100.0	100.0
	M00039	V$CREB_01	0.77	-336 - -324	32	0.0041	100.0	100.0	97.6	98.0	98.0	97.3
	M00033	V$P300_01	0.80	-24 - -18	32	0.0070						66.7
	M00023	V$HOX13_01	0.72	-47 - -37	32	0.0049						100.0
	M00191	V$ER_Q6	0.73	-11 - -1	31	0.0040						
	M00037	V$NFE2_01	0.79	-14 - -2	31	0.0036		100.0	100.0	100.0	87.5	78.3
	M00261	V$OLF1_01	0.76	-189 - -183	30	0.0081		100.0	100.0	100.0	100.0	100.0
	M00249	V$CHOP_01	0.77	-235 - -225	30	0.0050	47.6	42.0	41.9	40.8	41.7	41.2
	M00192	V$GR_Q6	0.77	-322 - -312	30	0.0052	100.0	94.8	94.8	95.1	95.1	96.1
	M00105	V$CDPCR3_01	0.75	-38 - -21	30	0.0028		100.0	100.0	100.0	100.0	100.0
	M00088	V$IK3_01	0.79	-147 - -137	30	0.0043		80.0	80.0	80.0	80.0	80.0
	M00183	V$MYB_Q6	0.82	-13 - -7	29	0.0072						
o	M00133	V$TST1_01	0.86	-284 - -273	29	0.0291						100.0
	M00086	V$IK1_01	0.77	-334 - -325	29	0.0056		100.0	100.0	100.0	100.0	100.0
o	M00215	V$SRF_C	0.75	-147 - -137	28	0.0117	100.0	100.0	100.0	100.0	98.0	98.6
	M00179	V$CREBP1_Q2	0.75	-270 - -263	28	0.0051						
	M00174	V$AP1_Q6	0.76	-9 - -5	28	0.0113						
	M00236	V$ARNT_01	0.75	-97 - -89	27	0.0047		100.0	100.0	100.0	100.0	100.0
	M00201	V$CEBP_C	0.80	-167 - -153	27	0.0026	93.3	87.6	87.2	85.8	86.7	86.1
	M00104	V$CDPCR1_01	0.77	63 - 73	27	0.0053		100.0	100.0	100.0	100.0	100.0
	M00069	V$YY1_02	0.75	-246 - -228	27	0.0025	100.0	84.2	86.4	89.7	90.0	91.4
o	M00240	V$NKX25_01	0.85	-231 - -219	26	0.0057						
o	M00238	V$BARBIE_01	0.78	-245 - -231	26	0.0107						
	M00217	V$USF_C	0.80	-223 - -218	26	0.0077	100.0	75.0	73.9	84.6	81.0	81.8
	M00173	V$AP1_Q2	0.79	-291 - -286	26	0.0076	80.0	84.0	84.8	81.5	83.5	83.7
	M00162	V$OCT1_06	0.80	-306 - -304	26	0.0174	50.0	95.0	95.8	95.6	95.3	95.6
	M00152	V$SRF_01	0.71	-42 - -28	26	0.0009						
	M00040	V$CREBP1_01	0.74	-36 - -28	26	0.0042		92.9	93.3	93.3	93.8	88.9
	M00062	V$IRF1_01	0.74	-349 - -339	25	0.0037	100.0	100.0	100.0	100.0	100.0	100.0
	M00011	V$EVI1_06	0.76	-163 - -154	25	0.0042						100.0
	M00279	V$MIF1_01	0.74	-19 - -6	24	0.0021						0.0
o	M00260	V$HLF_01	0.82	-116 - -103	24	0.0117		100.0	100.0	100.0	100.0	100.0
	M00251	V$XBP1_01	0.74	-35 - -24	24	0.0027						100.0
o	M00195	V$OCT1_Q6	0.79	-170 - -157	24	0.0093						
	M00138	V$OCT1_04	0.78	-347 - -339	24	0.0063	100.0	90.0	87.5	94.1	94.1	94.5
	M00007	V$ELK1_01	0.75	-17 - -13	24	0.0095		100.0	100.0	100.0	100.0	88.9
	M00190	V$CEBP_Q2	0.82	-76 - -66	23	0.0035						
	M00134	V$HNF4_01	0.76	-178 - -171	23	0.0049	100.0	100.0	100.0	100.0	100.0	100.0
	M00124	V$PBX1_02	0.74	-86 - -71	23	0.0020						
o	M00101	V$CDXA_02	0.98	-134 - -114	23	0.0258						
	M00082	V$EVI1_05	0.77	-39 - -34	23	0.0057						100.0
o	M00241	V$NKX25_02	0.83	-234 - -224	22	0.0260						
o	M00203	V$GATA_C	0.83	-168 - -158	22	0.0183	75.0	77.8	76.6	81.9	84.4	84.4
o	M00129	V$HFH1_01	0.78	-289 - -283	22	0.0104						
	M00109	V$CEBPB_01	0.81	-138 - -134	22	0.0096	100.0	94.1	91.7	88.9	88.9	89.2
	M00041	V$CREBP1CJUN_01	0.79	-336 - -327	21	0.0032	50.0	56.5	56.0	56.0	55.6	41.2
o	M00148	V$SRY_01	0.90	-336 - -333	20	0.0487		57.1	57.1	57.1	57.1	62.5
	M00107	V$E2_01	0.74	-274 - -266	20	0.0034		100.0	100.0	100.0	100.0	100.0
	M00042	V$SOX5_01	0.79	-298 - -294	20	0.0075		50.0	50.0	50.0	50.0	50.0
	M00157	V$RORA2_01	0.77	-36 - -24	19	0.0019						
	M00059	V$YY1_01	0.77	-294 - -292	19	0.0102	100.0	100.0	100.0	100.0	100.0	97.1
	M00258	V$ISRE_01	0.73	-218 - -209	18	0.0027						100.0
	M00233	V$MEF2_04	0.72	-43 - -31	18	0.0011	0.0	0.0	0.0	50.0	50.0	68.4
	M00159	V$CEBP_01	0.87	-248 - -243	18	0.0048	71.4	58.0	57.1	57.5	59.8	60.0

the maximum limit of the window width to 9 bp, such global overrepresentation cannot be detected. Although it is possible to deal with a wider local overrepresentation by setting the limit higher, we still have the problem that it requires considerable time for the trial and for estimating the background rate.

Acknowledgements

We wish to thank all the people who freely provided their data. Dr Prestridge kindly provided a list of non-promoter sequences. EPD R.50 was made available by Dr Bucher and Dr Trifonov. TRANSFAC was made by Dr Wingender *et al.*

Table 2. *Continued*

	ACCESS	FACTOR	CUT-OFF	Pref. reg.	#pro	background	Q1	Q1-2	Q1-3	Q1-4	Q1-5	Q1-6
	M00106	V$CDPCR3HD_01	0.82	-223 - -217	18	0.0050		100.0	100.0	100.0	100.0	100.0
o	M00100	V$CDXA_01	0.91	-257 - -252	18	0.0315						
	M00221	V$SREBP1_02	0.72	-245 - -240	17	0.0038	100.0	75.0	75.0	75.0	75.0	75.0
o	M00206	V$HNF1_C	0.79	-323 - -315	17	0.0101	77.8	85.7	87.5	94.1	94.7	94.7
	M00078	V$EVI1_01	0.72	-291 - -278	17	0.0016						100.0
o	M00212	V$POLY_C	0.74	-334 - -325	16	0.0058						
	M00188	V$AP1_Q4	0.79	-11 - -7	16	0.0066						
o	M00186	V$SRF_Q6	0.76	-348 - -342	16	0.0074						
	M00123	V$MYCMAX_02	0.78	-193 - -190	16	0.0062	100.0	80.0	72.7	76.9	76.9	79.2
	M00054	V$NFKAPPAB_01	0.78	88 - 91	16	0.0052	100.0	100.0	100.0	100.0	100.0	100.0
	M00224	V$STAT1_01	0.72	-84 - -72	15	0.0014			100.0	100.0	100.0	100.0
	M00199	V$AP1_C	0.77	-33 - -31	15	0.0078	70.0	78.6	80.6	87.2	87.6	87.8
	M00194	V$NFKB_Q6	0.78	-259 - -255	15	0.0049						
	M00205	V$GRE_C	0.77	-250 - -245	14	0.0042	80.0	86.4	86.4	87.1	87.1	86.2
	M00146	V$HSF1_01	0.76	-91 - -89	14	0.0069		100.0	100.0	100.0	100.0	100.0
	M00073	V$DELTAEF1_01	0.81	-305 - -303	14	0.0067						
	M00036	V$VJUN_01	0.73	-59 - -47	14	0.0010		0.0	0.0	0.0	0.0	0.0
	M00220	V$SREBP1_01	0.76	-342 - -336	13	0.0041	50.0	25.0	25.0	25.0	25.0	25.0
o	M00145	V$BRN2_01	0.84	-134 - -118	13	0.0138						
	M00119	V$MAX_01	0.73	-195 - -189	13	0.0022	100.0	100.0	100.0	100.0	100.0	100.0
	M00006	V$MEF2_01	0.73	-41 - -25	13	0.0009	100.0	100.0	100.0	100.0	100.0	68.2
	M00161	V$OCT1_05	0.85	-325 - -304	12	0.0004	0.0	85.0	87.5	76.5	72.9	74.7
	M00156	V$RORA1_01	0.76	-93 - -90	12	0.0052	100.0	100.0	100.0	100.0	100.0	100.0
	M00097	V$PAX6_01	0.75	-250 - -240	12	0.0017						
	M00256	V$NRSF_01	0.71	-10 - 2	11	0.0007						
	M00242	V$PPARA_01	0.73	-122 - -113	11	0.0015		100.0	100.0	100.0	100.0	100.0
	M00211	V$PADS_C	0.82	-207 - -205	11	0.0044						
	M00144	V$PAX5_02	0.74	-207 - -205	11	0.0050				100.0	100.0	100.0
o	M00130	V$HFH2_01	0.89	-168 - -158	11	0.0307						
	M00068	V$HEN1_01	0.73	-347 - -343	11	0.0027		100.0	100.0	100.0	100.0	100.0
o	M00232	V$MEF2_03	0.75	-218 - -214	10	0.0074	100.0	100.0	100.0	100.0	100.0	100.0
	M00225	V$STAT3_01	0.71	-339 - -327	10	0.0010			100.0	100.0	100.0	100.0
	M00223	V$STAT_01	0.79	-288 - -286	10	0.0062		100.0	100.0	100.0	100.0	100.0
	M00200	V$CAAT_C	0.71	-162 - -160	10	0.0045						
o	M00131	V$HNF3B_01	0.85	-215 - -212	10	0.0142	100.0	100.0	40.0	40.0	40.0	40.0
o	M00096	V$PBX1_01	0.89	-182 - -176	10	0.0361						
	M00087	V$IK2_01	0.85	-254 - -252	10	0.0042		75.0	75.0	75.0	75.0	75.0
o	M00228	V$VBP_01	0.81	-62 - -58	9	0.0149		100.0	100.0	100.0	100.0	100.0
	M00222	V$TH1E47_01	0.79	-237 - -235	9	0.0048		66.7	66.7	66.7	66.7	57.1
	M00150	V$BRACH_01	0.72	-221 - -184	9	0.0002						
o	M00137	V$OCT1_03	0.84	-330 - -327	9	0.0222	100.0	65.0	62.5	54.4	47.1	48.4
	M00116	V$CEBPA_01	0.80	-136 - -134	9	0.0062	95.2	78.0	78.1	76.7	78.0	78.5
	M00063	V$IRF2_01	0.70	-77 - -75	9	0.0048			100.0	100.0	100.0	100.0
	M00024	V$E2F_01	0.73	-57 - -54	9	0.0023				100.0	100.0	100.0
	M00135	V$OCT1_01	0.73	-322 - -320	8	0.0034	100.0	95.0	95.8	94.1	92.9	93.4
	M00117	V$CEBPB_02	0.84	-136 - -134	8	0.0026	100.0	94.1	95.8	88.9	88.9	81.1
	M00103	V$CLOX_01	0.76	-274 - -266	8	0.0015						
	M00095	V$CDP_01	0.74	-180 - -176	8	0.0028		100.0	100.0	100.0	100.0	100.0
	M00070	V$TAL1BETAITF2_01	0.76	-239 - -236	8	0.0027		100.0	100.0	100.0	100.0	100.0
o	M00026	V$RSRFC4_01	0.79	-166 - -154	8	0.0076						100.0
	M00210	V$OCT_C	0.76	-212 - -210	7	0.0031	100.0	100.0	100.0	88.9	85.6	88.3
	M00079	V$EVI1_02	0.77	-188 - -185	7	0.0027						100.0
	M00065	V$TAL1BETAE47_01	0.77	-318 - -314	7	0.0020		100.0	100.0	100.0	100.0	100.0
o	M00045	V$E4BP4_01	0.76	-194 - -192	7	0.0068		100.0	100.0	100.0	100.0	100.0
o	M00128	V$GATA1_04	0.81	-12 - -10	6	0.0198	96.4	90.0	90.5	89.2	89.5	89.5
	M00035	V$VMAF_01	0.76	-11 - -7	6	0.0015		100.0	100.0	100.0	100.0	100.0
o	M00268	V$XFD2_01	0.83	-64 - -62	5	0.0121						
	M00250	V$GFI1_01	0.75	-235 - -233	5	0.0023						
o	M00231	V$MEF2_02	0.77	-99 - -97	5	0.0077	100.0	100.0	100.0	100.0	100.0	100.0
	M00214	V$SEF1_C	0.70	-223 - -220	5	0.0014						100.0
	M00136	V$OCT1_02	0.77	-299 - -297	5	0.0037	100.0	90.0	87.5	88.2	88.2	89.0
o	M00132	V$HNF1_01	0.86	-70 - -56	5	0.0067	55.6	57.1	62.5	73.5	68.4	68.4
	M00248	V$OCT1_07	0.79	-237 - -235	4	0.0020	0.0	87.0	85.2	90.5	87.0	88.3
	M00102	V$CDP_02	0.81	-336 - -325	4	0.0004		100.0	100.0	100.0	100.0	100.0
	M00034	V$P53_01	0.70	-68 - -55	4	0.0001			100.0	100.0	100.0	75.0
o	M00267	V$XFD1_01	0.85	-350 - -350	0	0.0148						100.0
o	M00081	V$EVI1_04	0.89	-350 - -350	0	0.0168						100.0

Special thanks to Dr Todd Taylor and Dr Wayne Dawson for commenting on the original manuscript. We are grateful to Dr Michael Q.Zhang at Cold Spring Harbor Laboratory and Prof. Nikolay A.Kolchanov at the Russian Academy of Sciences for helpful discussions. We would also like to thank the anonymous referees for suggesting many and fruitful comments. This work is partially supported by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports, and Culture, Japan.

Appendix A. Method for calculating the binding score at each position on DNA

We calculate and normalize the binding scores between TFs and DNA using PWM according to Bucher (1990). From the frequency matrix, the PWM is made according to the calculation: $W_{b,i} = \ln(p_i(b) + a)$ where b refers to the base ($b \in \{A, T, G, C\}$) at position i of the input sequence, $p_i(b)$ is the probability of each base b at each position i, and a is a smoothing parameter ($\equiv 0.01$).

For the input sequence, the binding score is calculated: $x = \sum_{i=1}^{L} W_{b,i}$ where L is the consensus length of the motif. To normalize the score, the hypothetical maximum score and minimum score are calculated: $x_{max} = \sum_{i=1}^{L} \max_b W_{b,i}$, $x_{min} = \sum_{i=1}^{L} \min_b W_{b,i}$.

Thus the score for the input sequence is normalized: match $= \frac{x - x_{\min}}{x_{\max} - x_{\min}}$.

Appendix B. Background estimation

The TF's binding probability of having at least one binding site within w base pairs [$\equiv p(w)$] is estimated on non-promoter sequences by counting binding sites within each window at every position. Let us suppose that k sequences are randomly given. The expected number of sequences having at least one binding site within w base pairs is $k \cdot p(w)$, and its standard deviation is $\sqrt{k \cdot p(w) \cdot (1 - p(w))}$, where binomial distribution is used. We assume here that promoters are independent of each other. Since we used non-redundant promoters in EPD, we applied this assumption. However, if some promoters are dependent on each other, e.g. when we use closely related promoters collected from the same tissue, binomial distribution will not hold; we must use better models reflecting such dependencies.

References

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Bucher,P. and Trifonov,E.N. (1986) Compilation and analysis of eukaryotic Pol II promoter sequences. *Nucleic Acids. Res.*, **14**, 10009–10026.

Chen,Q.K., Hertz,J.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

Frech,K., Herrmann,G. and Werner,T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.*, **21**, 1655–1664.

Frech,K., Quandt,K. and Werner,T. (1997) Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.*, **13**, 89–97.

Fujibuchi,W. and Kanehisa,M. (1997) Prediction of gene expression specificity by promoter sequence patterns. *DNA Res.*, **4**, 81–90.

Goodrich,J.A., Schwartz,M.L. and McClure,W.R. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). *Nucleic Acids Res.*, **18**, 4993–5000.

Heinemeyer,T. *et al.* (1998) Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucleic Acids Res.*, **26**, 364–370.

Hertz,G.Z. and Stormo,G.D. (1995) Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In Lim,H.A. and Cantor,C.R. (eds), *Bioinformatics and Genome Research: Proceedings of the Third International Conference on Bioinformatics and Genome Research*. World Scientific, Singapore, pp. 201–216.

Hertz,G.Z., Hartzell,G.W. and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Laurence,C.E., Atschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,F. and Wootton,J.C. (1993). Detecting subtle sequence analysis—a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 108–214.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.

Prestridge,D.S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

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

Tsunoda,T. and Takagi,T. (1998) Automatic extraction of position specific co-occurrence of transcription factor bindings on promoters. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing '98*. World Scientific, Singapore, pp. 252—263.

Wolfertstetter,F., Frech,K., Herrmann,G. and Werner,T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12**, 71–80.