Algorithm Determining All Cut-off Values of TF-DNA Binding Score Calculated Using PWMs in TRANSFAC

Tatsuhiko Tsunoda **Toshihisa Takagi**
`tatsu@ims.u-tokyo.ac.jp` `takagi@ims.u-tokyo.ac.jp`
Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: transcription factor, transcription regulation, positional weight matrix, cut-off score, transcription factor binding sites

1 Introduction

Precise analysis of the genetic network, gene function, and transcription regulation requires accurate prediction of transcription factor (TF) bindability on DNA. A typical method calculates TF binding score for each site using positional weight matrix (PWM), and pick up candidates which exceed a given cut-off. However, such cut-offs were not determined for all TFs since we had no robust criteria. Although we define the optimum cut-off value that can correctly discriminate functional sites from background sequences, in general, functional sites are not given explicitly, since we poorly know where on DNA each TF binds. Positive and negative instances of binding sites are very difficult to collect exhaustively. Even if we could do it, we can not determine each cutoff value uniquely since they still have a freedom to solve. Thus we decided to fully use the assumption that such functional sites are conserved at certain region in DNA[1] since transcription regulation is constructed on 3D biochemical apparatus of DNA-TF interaction. They can be mined by estimating the local over-representation (LOR). Detecting multiple LORs independent of TF and promoter structure, our algorithm managed to determine the cut-off values for all (205) PWM of TF in TRANSFAC.

2 Method and Results

For detecting the conserved functional sites, we newly introduce a generalized LOR (see also Figure 1):

$$O_g = \frac{Detected \ \# \ of \ signals \ within \ a \ window - Average \ background \ for \ the \ window \ size}{Standard \ deviation \ of \ the \ background},$$

where each depends on factor, cut-off for TF-binding score, and window size. The number of signals and O_g depends also on position in promoters. O_g shows the significance of the detected number of promoters that bind the TF compared with the random fluctuation. However, the full set of promoters ($\equiv \mathcal{S}$) consists of two types of promoters: promoters in which the TF functional sites are conserved in the preferred region during evolution ($\equiv \mathcal{S}_a$), and others ($\equiv \mathcal{S}_b$). To discriminate $\mathcal{S}_a$ from $\mathcal{S}_b$, we must determine the cut-off beforehand. That is, the processes of determining the cut-off and discriminating $\mathcal{S}_a$ from $\mathcal{S}_b$ are mutually dependent.

Using an initial th, we can calculate TF binding sites in promoters. If we find many promoters that have TF binding sites within the same window, they will be functional. We separate $\mathcal{S}$ into two subsets: $\mathcal{S}_a$, in which promoters have TF binding sites within the window, e.g. TATA-containing promoters, and $\mathcal{S}_b$ which do not. Here, we check whether $\mathcal{S}_b$ does not have LOR within the window even if TF binding sites are re-estimated at any hypothetical threshold th' lower than th. In $\mathcal{S}_b$, if by temporarily lowering the cut-off value, statistically significant LOR is detected, then there is evidence that it consists of functional sites. However, this is opposed to the definition of $\mathcal{S}_b$; it means that some promoters that should be classified into $\mathcal{S}_a$ are mis-classified into $\mathcal{S}_b$. Thus we must reduce th with some step, separate $\mathcal{S}$ into $\mathcal{S}_a$ and $\mathcal{S}_b$, and recheck. We repeat this until $\mathcal{S}_b$ does not have LOR within the window even if TF binding sites are re-estimated at any hypothetical threshold th' lower than th.

Since preferred regions can be found multiply and anywhere in the promoters, we consider the cut-off to be optimum if it satisfies two following conditions: (1) Anywhere in the promoters, $\mathcal{S}_b$ does not have LOR within the window even if TF binding sites are re-estimated at any hypothetical threshold th' lower than th, and (2) Maximum cut-off that satisfies (1).

We used 205 vertebrate TFs from the database TRANSFAC Ver.3.4 [3], and EPD R.50 [2] for promoter sequences. Our final set consisted of non-redundant 433 promoters for which the region -349 – +100 bp of the TSS had been determined. From GenBank, we extracted sequences totaling 664,505 bp (1,329,010 bases) according to the list of non-promoters from Dr. Prestridge at Minnesota University [4]. The estimated cut-off value, background rate using the cut-off value of each TF is shown in Table 1.

```
Algorithm of cut-off determination
INPUT: $\mathcal{S}, PWM_f$
OUTPUT: optimum cut-off value for $f$
begin
  align $\mathcal{S}$ with TSS; $th := 1.0$;
  for $x := x_{min}$ to $x_{max}$ do
    for $w := w_{min}$ to $w_{max}$ do
      begin
        repeat
          Search signals of $f$ on $P_k$ in $\mathcal{S}$ using
          $PWM_f$ and $th$ within the window;
          Separate $\mathcal{S}$ into $\mathcal{S}_a$ and $\mathcal{S}_b$;
          $th' := th$;
          repeat
            $th' := th' - step$;
            Search signals of $f$ on $P_k$ in $\mathcal{S}_b$ using
            $PWM_f$ and $th'$ within the window;
            Count $N(\ \mathcal{S}_b, f, th', x, w)$;
            Calculate $O_g(\mathcal{S}_b, f, th', x, w)$;
            if $O_g > O_c$ and $N > N_c$ then LOR is
            detected in $\mathcal{S}_b$;
          until $th' < 0$ or LOR is detected in $\mathcal{S}_b$
          if LOR is detected in $\mathcal{S}_b$
            then $th = th - step$;
        until LOR is not detected in $\mathcal{S}_b$
      end ;
end ;
```

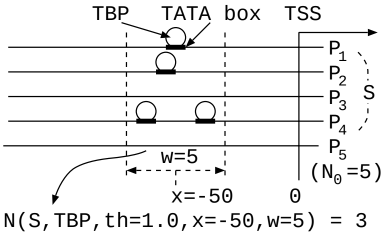

Figure 1. Promoter alignment, local window, and TF binding sites.

Table 1. Final results: estimated cut-off value and background rate.

ACCESS	FACTOR	CUT-OFF	Pref. reg.	#pro	background
M00189	V$AP2_Q6	0.78	-173 - -36	391	0.0269
M00008	V$SP1_01	0.78	-69 - -35	323	0.0297
M00252	V$TATA_01	0.77	-40 - -23	297	0.0065
M00255	V$GC_01	0.78	-74 - -45	292	0.0243
M00175	V$AP4_Q5	0.78	32 - 65	250	0.0175
M00253	V$CAP_01	0.87	-5 - 6	179	0.0226
M00254	V$CAAT_01	0.78	-105 - -70	174	0.0093
:	:	:	:	:	:

Cut-offs for all (205) TFs which have frequency matrices in TRANSFAC could be determined using our algorithm.

For detailed description of the algorithm and results, please see our full paper[5]. The cut-off values and transcription factor binding site predicting tool are also available at our WWW site[6]. This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science" from the Ministry of Education, Science, Sports, and Culture, Japan.

References

[1] Bucher, P., Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences, *J. Mol. Biol.*, 212(4):563–578, 1990.

[2] Bucher, P. and Trifonov, E.N., Compilation and analysis of eukaryotic POL II promoter sequences, *Nucleic Acids Res.*, 14(24):10009–10026, 1986.

[3] Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L. and Kolchanov, N.A., Databases on transcriptional regulation, *Nucleic Acids Res.*, 26(1):362–367, 1998.

[4] Prestridge, D.S., Predicting Pol II promoter sequences using transcription factor binding sites, *J. Mol. Biol.*, 249(5):923–932, 1995.

[5] Tsunoda, T. and Takagi, T.: *Bioinformatics*, 15(7/8):622–630, 1999.

[6] http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/TFBIND.