Detection of Preferentially Co-occurring Transcription Factor Binding Sites

Holger Klein, Martin Vingron
Department of Computational Molecular Biology
MPI for Molecular Genetics

April 15, 2005

1 Abstract

In this work we first discuss approaches for detecting pairs of transcription factor binding sites (TFBSs) which tend to co-occur in regulatory regions of vertebrate promoters and thus are potentially synergistic. We search for putative TFBSs in a set of human sequences from the Eukaryotic Promoter Database using TRANSFAC matrices. TFBSs are then grouped to avoid the detection of putative synergistic pairs solely because of similarity of corresponding position weight matrices. One approach for counting pairs of TFBSs in a given window and the subsequent calculation of a pair-score for the relative overrepresentation of a given pair of two TFBSs is presented, alternative ways are outlined. Some high scoring pairs of known interacting transcription factors are are found with the described method. Necessary extensions are discussed.

2 Introduction

Transcriptional regulation is controlled by protein complexes of transcription factors binding to specific DNA sequences around and upstream of the transcription start site (TSS). Because of this binding sites for transcription factors (TFBSs) contained in these protein complexes often occur in proximity to each other. The TFBSs can be found organized in cis-regulatory modules or clusters. For metazoans a cis-regulatory module typically consists of up to ten binding sites for at least three different sequence-specific transcription factors stretched over roughly 500bp [5]. These modules can function to direct complex spatial or temporal expression patterns.

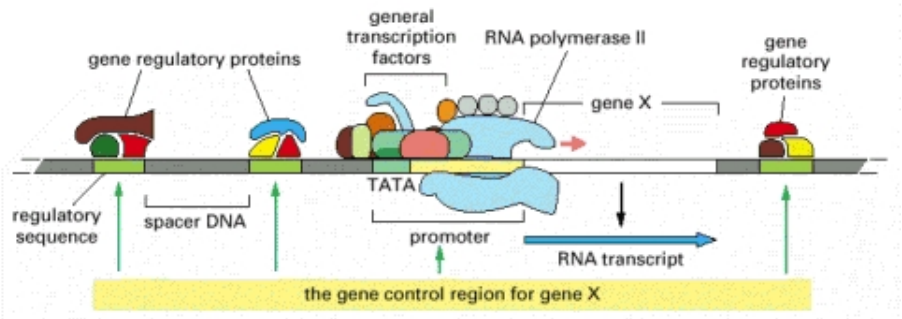

Figure 1: Gene control region of a typical eukaryotic gene (picture taken from [1]).

The goal of this project is to predict putative synergistic or antagonistic transcription factors based on an analysis of the co-occurrence of their binding sites in known regulatory regions.

3 Methods

3.1 Scanning for and merging of TFBSs

TRANSFAC is a database containing descriptions of cis-regulatory DNA elements. These are derived from known transcription factor binding sites and are represented by gap-less profiles or position weight matrices (PWMs) [7].
TRANSCompel is a database containing a set of of known interacting transcription factors [3], which is used to validate putative synergistic pairs of transcription factors.
Using the methods of Rahmann et al. [8] we scan the set of sequences to be searched for overrepresented co-occurring pairs of TFBSs for the set of all vertebrate matrices present in TRANSFAC.

In some cases PWMs belonging to different factors are very similar and TFBSs for such factors are predicted at the same position. As a consequence predictions of pairs of such co-occurring TFBSs are overrepresented. To reduce the influence of PWM similarity, TFBSs are merged into and counted as one if they belong to the same factor or PWM class. We consider three different sets F_1, F_2 and F_3 of PWM families:

- F_1 based on the binding factor classes as defined in TRANSFAC, [10] resulting in 220 groups of matrices.
- F_2 based on the method described by Schones et al. [12], grouping PWMs into families based on the similarity of matrices, resulting in 145 matrix families.
- F_3 based on a grouping into PWM families based on a distance measure and subsequent clustering [10], resulting in 45 different matrix families.

In the sequel we assign labels to predicted TFBSs that refer to either a single PWM hit or merged hits of PWMs belonging to the same family.

3.2 Counting and calculation of a co-occurrence score

Adapted from the procedure described by Rateitschak et al. [9] we count the combinations of TFBSs within a given sequence window, sliding over each sequence in the dataset. To further reduce the influence of matrix similarity we also take into account the possibility of a minimum distance of two TFBSs (from the center of one hit to the center of the other) to be counted as a pair. Strand orientation and pair order relative to TSS are disregarded.
While in [9] only one occurrence of a pair of TFBSs per sequence is considered, in this study we count more than one occurrence per sequence.
Using a sliding window can lead to counting a pair of TFBSs which has already been counted in a previous window a second time. To reduce the number of times this happens, we shift the window of interest on the sequence by different step sizes such that the overlapping parts of two neighboring windows are small relative to the window size.
After counting the pairs, we calculate a score S_{ij} for each pair (i,j) of TFBSs.

$$S_{ij} := log \frac{m_{ij}}{\pi_i \pi_j} \tag{1}$$

with

$$m_{ij} = \frac{f_{ij}}{\sum_{k,l} f_{kl}}$$

and

$$\pi_i = \frac{\sum_k f_{ki}}{\sum_{k,l} f_{kl}}$$

where f_{ij} is the number of counted pairs of $TFBS_i$ and $TFBS_j$, π_i/π_j the number of pairs of $TFBS_i/TFBS_j$ with arbitrary partners and $\sum_{k,l} f_{kl}$ the total number of pairs.
A positive score S_{ij} stands for an overrepresentation of a pair (i,j) of TFBSs, while a negative score signifies a smaller than expected number of occurrences of a pair (i,j) in the dataset.

3.3 Dataset

We extract 1796 non-redundant human promoter sequences from the Eukaryotic Promoter Database (EPD, rel. 81, [11]), a representative set of sequences not sharing more than 50% sequence identity with each other. Used sequences range from position -500 to +100 relative to the experimentally verified transcription start site.

4 Results

4.1 Calculation of co-occurrence scores S_{ij}

We calculate co-occurrence scores for the three described methods of merging the TFBS hits. We recalculate the scores for three minimum distances between the centers (0bp, 10bp, and 25bp) of two TFBS hits and three window sizes (50bp, 100bp, 200bp) to assess the influence of the parameters. We extract high scoring pairs for the different calculations.
In figures 2 and 3 we show the results for the TFBS annotations based on the set F_3 of PWM families, a window size of 100bp, and minimum distances between the centers of TFBS hits of 0bp, 10bp, and 25bp. In the colormap yellow areas represent high scores, while blue areas represent low scores. The diagonal shows the scores for homotypic pairs of TFBSs. The lowest and the rightmost lines of the map show the scores for TFBSs occurring in a window without any partner. The matrix is symmetrical, because at the moment the implementation of the score calculation does not distinguish between different orientations of a pair of TFBSs (e.g. a pair A-B is counted the same as a pair B-A). Figure 3 shows a histogram of the scores for the described dataset.
Table 1 shows the transcription factor pairs which have a score $S_{ij} > 1.5$ calculated with formula 1 using different minimum distances and a window size of 100bp.

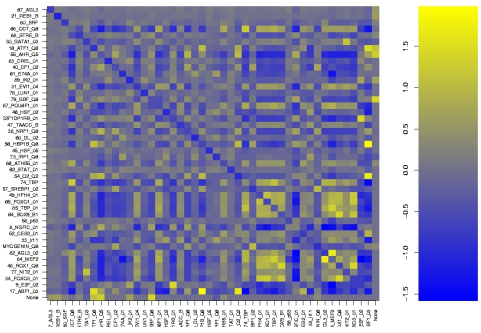

Figure 2: Score matrix with scores S_{ij} for EPD promoter set, scanned for TFBSs with PWM family set F_3, window size 100bp, minimum distance between center of hits 10bp.

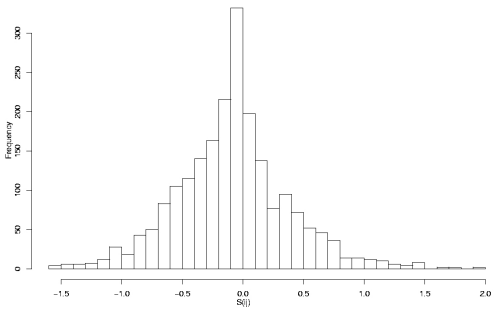

Figure 3: Histogram of scores S_{ij} calculated for EPD promoter set, scanned for TFBSs with PWM family set F_3, window size 100bp, minimum distance between center of hits 10bp.

In table 1 the influence of the minimum distance parameter can be seen. For the high scoring pairs the scores normally grow when a minimum distance is applied. In these cases this behaviour is expected, because of the influence of the total number of pairs and the number of pairs in m_{ij}, π_i, and π_j and the number of pairs with arbitrary partners in π_i and π_j. The total number of pairs is reduced to roughly 25% when comparing the results for a minimum distance of 0bp to a

pair		S_{ij} (min. dist. 0)	S_{ij} (min. dist. 10)	S_{ij} (min. dist. 25)
None	16_ATF1_Q6	1.7377	1.0991	0.7664
17_ABF1_03	54_O2_Q2	1.5737	1.6848	2.0350
17_ABF1_03	56_HBP1B_Q6	1.8066	1.9756	2.2992
84_MEF2	82_AGL3_02	1.6367	1.7337	1.9734
9_E2F_02	12_E2F1DP1RB_01	1.2930	1.4187	1.6935
16_ATF1_Q6	17_ABF1_03	1.3576	1.4994	1.8256
34_FOXO3_01	48_ROX1_Q6	1.3348	1.4172	1.6294
48_ROX1_Q6	64_SOX9_B1	1.3070	1.4060	1.6421

Table 1: High scoring pairs of TFBSs for EPD promoter set, scanned for TFBSs with PWM family set F_3, window size 100bp, and minimum distances 0, 10, and 25 bp.

minimum distance of 25bp, since at a window size of 100bp this means a reduction of sequence area searched for TFBSs of 50%. An exception is the pair "None - 16_ATF1_Q6", which can be explained by the relatively higher number of TFBSs occurring on their own in a window when imposing minimum distance constraints.
Comparing the list of highest scoring pairs to TRANSCompel one can find examples like E2F/E2F (co-occurrence score $S_{ij} = 1.4187$ at parameters: minimum distance 0, window size 100, 99.50% of scores in dataset smaller). Most known interacting pairs can be found at small to medium positive co-occurrence scores though (e.g. CEBP/Stat: $S_{ij} = 0.08$, 70.88% of scores in dataset smaller).
Some of the factors from the TRANSCOMPEL list cannot be connected to PWM families with certainty. One reason is due to ambiguous relations between factors and their position weight matrix clusters. Another reason is that even in a case of a unique relationship between a factor and a PWM, the PWM might be grouped with PWMs from other factors due to matrix similarity.

5 Discussion and further work

In this work we present a method of calculating a co-occurrence score for putative synergistic pairs of transcription factors. A few known interacting pairs of TFBSs are found at high scores, more at medium to low positive scores. Several parameters have an impact on the actual scores, which still has to be evaluated. A more thorough assessment of the results is needed. This can be done with a well defined set of interacting pairs of TFs taken from TRANSCompel, TRANSFAC and other sources, since many more interactions are known than are included in these databases. Another possibility of testing the approach is to take a set of well known regulatory regions of genes expressed in a certain tissue (for example liver, [4]) and check if the expected pairs of TFs are assigned high scores.
The choice of PWM grouping is important. Having many PWMs in one family avoids counting pairs just based on similarity of the corresponding PWMs. The drawback is the ambiguity of the assignment of a PWM family to a TF, which makes it more difficult to validate the predicted synergistic pairs.
The influence of the other parameters, such as minimum distance of two TFBSs and window size for the counting procedure have to be thoroughly assessed, which is ongoing work. The preliminary impression is that most pairs of TFBS with high scores show up at high scores independent of the window size and the minimum distance criterium, although the rank of a pair might change.
To overcome the problem of counting the same pair twice the method has been enhanced. Now a pair of TFBSs is not counted, if excactly the same pair of TFBSs has been counted previously and if a pair of TFBSs belonging to the same transcription factors is present within the given window. Using this method we increment the pair count for two TFBSs if a combination is present within a window *once or more* and if this combination has not been counted within the same window.
Also a new method of calculating the expected number of pairs has been implemented. Similar to the approach used by Levy et al. [6] the labels of the TFBSs found on a sequence are permuted and the pairs of TFBSs are counted. This procedure is repeated several times and the average

number of pairs for every combination is calculated. Now a log-odds score of the counted number of pairs and the average number of pairs derived by the permutations can be calculated.
Regarding the sets of sequences that are to be analyzed, the next step is to look only at conserved non coding regions of the EPD dataset [2].

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell. 4th ed.* Garland Publishing, 2002.

[2] Christoph Dieterich, Steffen Grossmann, Andrea Tanzer, Stefan Ropcke, Peter Arndt, Peter Stadler, and Martin Vingron. Comparative promoter region analysis powered by CORG. *BMC Genomics*, 6(1):24, Feb 2005.

[3] Olga V Kel-Margoulis, Dmitri Tchekmenev, Alexander E Kel, Ellen Goessling, Klaus Hornischer, Birgit Lewicki-Potapov, and Edgar Wingender. Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol*, 3(1-2):145–71, 2003.

[4] W Krivan and WW Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*, 11(9):1559–66, Sep 2001.

[5] Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–51, Jul 2003.

[6] S Levy, S Hannenhalli, and C Workman. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, 17(10):871–7, Oct 2001.

[7] V Matys, E Fricke, R Geffers, E Gossling, M Haubrock, R Hehl, K Hornischer, D Karas, AE Kel, OV Kel-Margoulis, D-U Kloos, S Land, B Lewicki-Potapov, H Michael, R Munch, I Reuter, S Rotert, H Saxel, M Scheer, S Thiele, and E Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–8, Jan 2003.

[8] Sven Rahmann, Tobias Mueller, and Martin Vingron. On the power of profiles for transcription factor binding site detec. *Statistical Applications in Genetics and Molecular Biology*, 2(1):Art. 7, 2003.

[9] Katja Rateitschak, Tobias Mueller, and Martin Vingron. Annotating significant pairs of transcription factor binding sites in regulatory DNA. *In Silico Biol*, 4(3):0040, 8 2004.

[10] Stefan Roepcke. personal communication.

[11] Christoph D Schmid, Viviane Praz, Mauro Delorenzi, Rouaeda Perier, and Philipp Bucher. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res*, 32(Database issue):D82–5, Jan 2004.

[12] Dustin E. Schones, Pavel Sumazin, and Michael Q. Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, Aug 2004.