A Fast Algorithm for Analysing Regulatory Regions of Related Genes

Matthieu Defrance, Hélène Touzet, Stéphane Janot
UMR CNRS 8022, LIFL, Université des Sciences et Technologies de Lille
{defrance, touzet, janot}@lifl.fr

keywords: Regulation, Transcription factor binding sites, Human genome

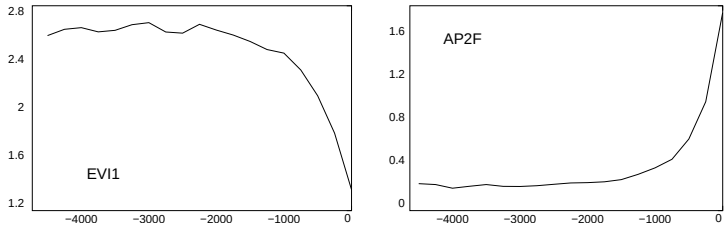
Introduction

The first step toward the modelling of a transcriptional genetic network is to understand which are the transcription factors that are involved in the process. It is possible to represent the binding site of a transcription factor on the DNA regulatory sequence with a position weight matrix, and so to locate potential binding sites on a regulatory region. However this approach requires a prior knowledge about the involved trancription factors and it is not suitable for long genomic sequences, due to the high number of false positive predictions.

It is more likely to infer regulation hypothesis for families of genes that are assumed to share regulatory mecanisms. It applies to functionally related genes or to clusters of genes that are derived from DNA array analysis. The aim of this work is to present a computational method that searchs for common regulatory elements in promoter regions of co-regulated genes. The algorithm developed here can deal with proximal as well as with distal sites and it is able to extract information from families of thousands base long sequences.

Mining the regulatory sequences

The core of the method consists in detecting regions where predicted binding sites show local over-representation. This idea is already present in [1]. The first task is to be able to properly define when a factor is over-represented. For that, we chose to rely on the observed distribution of predicted binding sites with Matinspector [6]. Those distribution are of course heterogen. Figure 1 shows two examples : distribution of EVI1 and AP2F factors for 5000 arbitrary Human promoters.



*The horizontal axis is the position on the promoter (*0 *is the transcription initiation site), and the vertical axis indicates the number of predicted binding sites per gene for 500 bp sliding windows.*

Figure 1. Two examples of distribution of predicted binding sites

It remains to extract significant regions. The method should not make any assumption about the length of the regulation region, its location on the DNA sequence, or the number of involved genes. A limited number of such candidate regions is detected by a linear greedy heuristic strategy. We select final regions amongst these candidates using a simple Chi square statistical test. For one transcription factor, this takes much less than one second. So it is possible to check a set of sequences against all matrices of the TRANSFAC transcription factor database [9] within a few seconds.

Examples

We demonstrate the relevance of our technique with two case studies in Human genome: 32 *muscle-specific genes* of the database [8] and 99 *NF-κB target genes* compiled from [2, 3, 4, 5].

For each gene, we recovered the Refseq accession number and we retrieved the corresponding promoter sequences from [7] (5000 bp upstream the initiation transcription site). We then searched exhaustively for all potential binding sites predicted with Matinspector from TRANSFAC vertebrate matrices. This brute output of Matinspector gives more than 90% of false positives. Finally, we looked for over-represented factors. The background distribution is here based of 5000 Human genes that were randomly picked up from [7].

muscle-specific genes		
Factor	Region	Score
TEAF	-318 -54	6.55^{-11}
MYOD	-2581 -2463	2.09^{-10}
MEF2	-82 -30	6.30^{-9}
TBPF	-42 -30	1.14^{-8}
MZF1	-305 -32	4.79^{-8}

NF-κB target genes		
Factor	Region	Score
NFKB	-230 -56	1.16^{-11}
TBPF	-67 -20	1.09^{-8}

Figure 2. Results for muscle-specific genes and TdkB target genes

In Figure 2 we report all predicted factors whose significance score is smaller than 10^{-7}. These factors are actually involved in the regulation mecanisms. Moreover the interesting regions (extracted from 5000 bp long sequences) are consistent with the observed binding sites. MYOD is a notable exception : it is predicted more than 2000 bp usptream the initiation site. It may be a sign of existence of alternative binding sites. Lastly, the method appears to be robust to uncertainty of data. For instance, for more than 20% of the muscle-specific genes regulatory elements are located in Intron 1 or Intron 2, and we restricted our inspection to upstream promoters.

Software availability

The method described here is implemented in C with a web interface. It is available from the authors upon request.

Acknowledgment

The authors are deeply grateful to Corinne Abbadie and Karo Gosselin (UMR CNRS 8117 - Institut de Biologie de Lille) for their valuable help providing us with the annotated data set of NF-κB target genes. They also thank Gilles Fay, Jacques van Helden and Bernard Vandenbunder for fruitful discussions.

References

[1] Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau Y., and De Moor,. B. (2003) TOUCAN: deciphering the cis-regulatory logic of coregulated genes, Nucleic Acids Research, 31(6), 1753-1764

[2] Karin, M. and Lin A. (2002), TdkB at the crossroads of life and death Nature Immunology 3 (3), 221-227

[3] Li X. et al. IKKalpha, IKKbeta and NEMO/IKKgamma are each required for the NF-kB mediated inflammatory response program JBC 277 (47), 45129-140

[4] Ohno H, Takimoto G, McKeithan TW. (1990), The candidate proto-oncogene bcl-3 is related to genes implicated in cell lineage determination and cell cycle control. Cell 23;60(6), 991-7

[5] Pahl, HL. (1999) Activators and target genes of Rel/TdkB transcription