Method for Recognition of Composite Regulatory Units in Promoters. Analysis of Cell-Cycle Control Genes

Alexander Kel[1] **Olga Kel-Margoulis**[1] **Edgar Wingender**[2]
kel@bionet.nsc.ru okel@bionet.nsc.ru ewi@gbf.de

[1] Institute of Cytology and Genetics SB RAN, pr. Lavrentyeva-10, 630090, Novosibirsk, Russia
[2] Gesellschaft für Biotechnologische Forshung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Keywords: site recognition, combinatorial regulation, E2F transcription factors, cell-cycle

1 Introduction

The particular expression pattern of genes is encoded in the structure of the transcription regulatory regions by specific combinations of transcription factor (TF) binding sites. Last years, several computational approaches have appeared addressing the problem of combinatorial regulation of transcription. Specific TF binding site combinations were identified for muscle-specific promoters [7, 3] for liver-enriched genes [6] and for yeast genes [1]. Recently, we have shown that search for specific combinations of two TF sites – composite elements – is a very effective tool for predicting gene expression patterns for immune-cell specific genes [4]. The drawback of all these approaches is their limitation to the only types of TF binding sites that is already known and sometimes quite poorly defined. The need of involving a wide spectrum of additional nucleotide signals in the analysis of promoter structure becomes clear now. New tools appeared, such as ModelGenerator that is able to invoke politracts and more complex feature in the analysis [2]. Here we present a new method for identifying combinatorial structures referred as TF composite units in promoters of similarly regulated genes.

2 Method and Results

2.1 Method for recognition of TF composite units

The method for recognition of TF composite units based on the SITEVIDEO approach developed earlier [5]. TF composite unit consists of a binding site for a known transcription factor Φ arranged with various flanking motifs $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ – potential targets for additional transcription factors. Such TF composite units could serve as combinatory targets for complexes of different transcription factors synergistically regulating gene transcription.

The method is designed by discriminating two set of sequences. In the first set (Y) sequences contain experimentally proven TF binding site for the considered factor Φ. A weight matrix designed for the target sites of this factor is used for aligning this set [4]. As the negative control set (N) we consider sequences matching this weight matrix as well, but yet known as containing no binding sites of the considered type. We search for short nucleotide motifs λ_i that appeared in the sequences of the set Y significantly more/less frequently then in the sequences of the set N. Means of SITEVIDEO system [5] are used for performing effective search through the space of all possible motifs in all possible windows. The significance is tested using many statistical criteria [5] integrated in a generalized criteria – utility value U. Motifs with the maximal utility are selected. Out of them, finally, we choose a limited set of uncorrelated motifs and use it for constructing a linear discriminating function.

Table 1: Frequency of potential E2F composite units

N	Set	Frequency (sites per 1.000bp)
1	C-gene promoters	3.298
2	EPD promoters	0.371
3	Exons	0.038

2.2 Search for E2F composite units in cell-cycle genes

Nowadays, one of the high priority field is the study of the molecular mechanisms regulating cell proliferation, commitment, differentiation and apoptosis. Understanding of these mechanisms will help to discover means for the treatment of cancer, developmental and immune diseases. Therefor analysis of transcriptional regulation of specific sets of genes activating during different stages of cell cycle is of high interest now. Using the methods developed in the present work we revealed E2F composite units in promoters of cell-cycle genes with maximal level of expression at the G1/S transition and in the S-phase of cell cycle (C-genes). The potential E2F composite unites exhibit extremely high frequency in promoters of C-genes in comparison with other promoters and exon sequences (see Table 1).

Search for E2F composite units in blind genomic sequences provides possibility to identify new cell-cycle genes playing important role in the complex network of gene regulation controlling cell proliferation, differentiation and apoptosis. Tools are available at
http://compel.bionet.nsc.ru/FunSite/SiteScan.html.

References

[1] Brazma, A., Vilo, J., and Ukkonen, E., Finding transcription factor binding site combinations in the yeast genome, *Proc. of the German Conference on Bioinformatics GCB'97*, Kloster Irsee, Bavaria, Sept. 22–24, 1997 (H.W.Mewes and D.Frishman eds.), 57–60, 1997.

[2] Frech, K., Danescu-Mayer, J., and Werner, T., A novel method to highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter, *J. Mol. Biol.*, 270:674–687, 1997.

[3] Frech, K., Quandt, K., and Werner, T., Muscle actin genes: A first step towards computational classification of tissue specific promoters, *In Silico Biology*, 1:0005, 1998. http://www.bioinfo.de/isb/1998/01/0005/

[4] Kel, A., Kel-Margoulis, O., Babenko, V., and Wingender, E., Recognition of NFATp/AP-1 Composite elements within genes induced upon the activation of immune cells, *J. Mol. Biol.*, 288:353–376, 1999.

[5] Kel, A.,E., Ponomarenko, M. P., Likhachev, E. A., Orlov, Yu. L., Ischenko, I. V., Milanesi, L., and Kolchanov, N. A., SITEVIDEO: A computer system for functional site analysis and recognition, Investigation of the human splice sites, *Comput. Appl. Biosci.*, 9:617–627, 1993.

[6] Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. and Pontoglio, M., Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome, *J. Mol. Biol.*, 266:231–245, 1997.

[7] Wasserman, W. W. and Fickett, J. W., Identification of regulatory regions which confer muscle-specific gene expression, *J. Mol. Biol.*, 278:167–181, 1998.