IN SILICO PREDICTION OF GENE REGULATION. COMPOSITE CLUSTERS OF TF BINDING SITES.

Alexander Kel, Olga Kel-Margoulis, Edgar Wingender
BIOBASE GmbH,Germany
http://www.biobase.de

Regulation of gene expression becomes the key problem of the era of "Functional Genomics". Now we know that genes in genomes of higher eukaryotic organisms are regulated mainly by the means of multiple regulatory proteins - transcription factors (TF), acting through specific regulatory sequences (TF binding sites) that are located usually in the proximity of the genes when constituting a promoter, or at more remote locations when being a part of an enhancer. Having available genomic sequences on one side and the massive though phenomenological gene expression data on the other side the challenge is to understand regulatory mechanisms of all and every gene in the genome by computer analysis of the gene regulatory sequences and by integrating this data with biological knowledge of the gene signal transduction, metabolic and physiological networks. Sophisticated computational regulatory sequence analysis tools that employ powerful statistical and machine learning algorithms driven by the rich databases that collect known biological facts enable us to make profound in silico predictions and formulate experimentally testable hypotheses. Such in silico driven experiments can greatly speed up the process of our understanding of the gene regulatory mechanisms and identification of new target genes. The understanding of how gene regulation mechanisms are encoded in the genomic regulatory sequences will give us a powerful means for deciphering causes of major human diseases. Functionally related genes involved in the same molecular-genetic, biochemical, or physiological process are often regulated coordinately by specific combinations of transcription factors. On the level of DNA, the blueprint of such common mechanisms of regulation may be seen as specific combinations of TF binding sites located in a close proximity to each other. We call such structures as "composite clusters" that could serve as good benchmarks for identification of regulatory regions in genomes. Last years, several computational approaches have appeared addressing the problem of combinatorial regulation of transcription. Specific TF binding site combinations were used for identification of muscle-specific promoters [1,2] for liver-enriched genes [3] and for yeast genes [4]. 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. We have demonstrated this approach for promoters of genes highly induced upon immune response [5]. Promoters of genes regulated during cell cycle could be recognized by combination of E2F binding sites with a dozen of oligonucleotide motifs [6]. A number of known examples of composite elements is collected in COMPEL database [7]. This data together with computationally predicted composite structure provide a key for annotation of regulatory regions in genomes. We have developed a method for revealing of composite clusters of cis-elements in promoters of eukaryotic genes that are functionally related or coexpressed. A software system "ClusterScan" have been created that enables: (i) to train system on representative samples of promoters to reveal cis-elements that tend to cluster; (ii) to train system on a number of samples of functionally related promoters to identify functionally coupled transcription factors; (iii) to provide tools for searching of this clusters in genomic sequences to identify and functionally characterize regulatory regions in genome. This tool was applied for regulatory annotation of human genome.

[1] Wasserman, W. W., Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. 278 (1998) 167-181.

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

[3] Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J. Mol. Biol. 266 (1997) 231-245.

[4] Brazma, A., Vilo, J. & Ukkonen, E. Finding Transcription Factor Binding Site Combinations in the Yeast Genome. In Proceedings of the German Conference on Bioinformatics GCB'97, Kloster Irsee, Bavaria, Sept. 22-24, 1997 (H.W.Mewes and D.Frishman eds.), (1997) 57-60.

[5] Kel, A., Kel-Margoulis, O., Babenko, V., Wingender, E. "Recognition of NFATp/AP-1 Composite Elements within Genes Induced upon the Activation of Immune Cells" J. Mol. Biol. 288 (1999) 353-376.

[6] Kel, A.E, Kel-Margoulis, O.V., Farnham, P.J., Bartley, S.M., Wingender, E., and Zhang, M.Q. Computer-assisted identification of cell cycle-related genes - new targets for E2F transcription factors. J. Mol. Biol. 309 (2001) 99 - 120.

[7] Kel-Margoulis, O.V., Romaschenko, A.G., Kolchanov, N.A., Wingender, E. and Kel, A.E. TRANSCompel: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res. 28 (2000) 311-315.