Identification of *cis*-Regulatory Elements and Their Combination in Mammalian Promoter Sequences

Igor Leykin, Ming-Chih Kao and Wing Hung Wong

Keywords: eukariotic promoters, regulatory elements.

Co-regulated genes may be regulated by common regulatory elements. Gene promoters are essential regulatory structures that regulate the initiation and level of transcription of a gene. Promoter is an integral part of the gene and often makes sense only in the context of its own gene, especially if important parts of the regulation are determined outside of the promoter. Computational prediction of eukariotic promoters solely from the nucleotide sequence is an attractive but difficult aspect of sequence analysis. Polymerase II promoters usually consist of multiple binding sites for transcription factors that must occur in a specific contex, apparently shared only by a small group of promoters[1].

Transcription factor binding sites are crucial promoter elements, and methods to detect such sites have been developed and used for more than a decade now. The majority of the known transcription factors recognize short DNA sequences of 5-15 bps in length with different degrees of internal variation. The individual binding of a transcription factor to a regulatory element is rarely sufficient to confer context-specific expression. Thus, the combination and orientation of the transcription factors is the crucial information, rather than the mere occurrence of several binding sites[2]. Cooperation between multiple factors interacting at multiple sites appears to be essential for gene regulation, but the biochemical rules governing these interactions remain largely unknown[3].

Our group is involved in extensive analyses of large scale microarray expression data from several clinical experiments as well as from experimental models in mice. All genes and ESTs can be clustered

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Av., Boston, MA 02115. E-mail: ileykin@hsph.harvard.edu

based on their expression pattern. For each cluster containing potentially co-regulated genes the functional organization of promoter sequences can be analyzed. One kilobase of upstream regions plus 50 bp of downstream regions can be substructed either directly from human genome assembly (for human experiments) or using The TIGR Orthologous Gene Alignment (TOGA) database[4] (for non-human data) and then analyzed for the common organizational framework of the transcription factor binding sites (GEMS Launcher, Genomatix[5]). The detected organizational framework will be common to all or user-defined percentage of the analyzed promoters. Individual promoters may contain additional functional transcription factor binding sites not included in the framework. The framework model should be sufficient to find the promoter, but is not a complete description of all functional elements. A model derived from such results can be used to scan databases for additional sequences that show a similar functional organization even in the absence of direct sequence similarity.

We have applied such approach to gene expression data from several projects such as Molecular Determinates of Breast Cancer and Identification of Circadian Clock Genes. We have identified several unique combination of multiple individual transcription factor binding sites in the clusters of similarly regulated genes. This method will allow us to characterize regulation of the gene expression using clinical and experimental data.

References:

1. Werner T. Models for prediction and recognition of eukariotic promoters. Mammalian Genome 1999, 10:168-175
2. Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analyses approach. J Mol Biol 2000, 297:599-606
3. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 1998, 278:167-81

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Av., Boston, MA 02115. E-mail: ileykin@hsph.harvard.edu

4. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva K, Antonesku V, Cho J, Parvizi P, Cheung F, Quackenbush J. Genome Biology 2001, 2(11):software0002.1-0002.4

5. Werner, T. Models for prediction and recognition of eukaryotic promoters. Mammalian Genome 1999, 10:168-175

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Av., Boston, MA 02115. E-mail: ileykin@hsph.harvard.edu