WHICH TRANSCRIPTION REGULATORY PATTERNS ARE MOST REPRESENTED IN HUMAN PROMOTERS

Vladimir B Bajic
BioDiscovery Group, Laboratories for Information Technology, Singapore
bajicv@lit.org.sg

Promoters of genes are the major regulatory regions for gene activation. They determine and modulate the gene's expression pattern based on specific combinations of transcription factor binding sites (TFBSs). Some of the experimentally verified functional combinations of TFBSs are deposited in the TRANSCompel database (Biobase, Germany). Diversity of promoter regions of eukaryotes make recognition of promoter regions and, more generally gene start, quite a difficult problem since in different promoters:

a/ TFBSs do not appear in the same combinations
b/ if they appear, their order differs from promoter to promoter
c/ mutual distances of TFBSs, as well as their distances from the transcription start site differ, from case to case
d/ there are no combinations of TFBSs that are common for a majority of promoters
e/ there are several thousands TFBSs described and collected.

The individual character of eukaryotic promoters is one of the greatest sources of difficulties to be overcome in designing efficient promoter and gene start recognition systems.

For this reason we investigated statistically the most overrepresented regulatory patterns in a large number of human gene promoters. We analyzed several thousands promoter regions, covering segments [-2000,+700] relative to the presumed transcription start sites of human genes. These extended promoter regions were extracted by the FIE program (http://sdmc.krdl.org.sg/FIE/). Statistically the most significant putative regulatory patterns, represented as combinations of TFBSs or oligonucleotides, are filtered out. The search for TFBSs was based on the Match program and TRANSFAC database of Biobase, Germany. The significance of individual TFBSs or their combinations, as well as of olugonucleotides and their combinations, was derived by analyzing large sets of exon and 3'UTR sequences in addition to promoter region sequences. This is the most comprehensive analysis of this type up to date. Some of the patterns found were already used in promoter and gene start recognition tools, such as Dragon Promoter Finder (http://sdmc.krdl.org.sg/promoter/) and Dragon Gene Start Finder
(http://sdmc.lit.org.sg/promoter/genestart0_1/genestart.htm).

These results show the relative significance of individual TFBSs and oligonucleotides, but particularly their combinations of two, three or more

patterns, which all together are characteristics of promoter and gene start regions. Such statistically important features could serve as a guideline for deciphering biologically significant patterns to support gene prediction, gene function inference, gene annotation, as well as to infer layers and interconnections of entities in complex regulatory networks.