Novel Computational Method for Human *Cis* Regulatory Elements Prediction

Alberto Ambesi-Impiombato [1,2] and Diego Di Bernardo [1]

[1] Telethon Institute of Genetics and Medicine, via P. Castellino 111,
80131 Naples - Italy
dibernardo@tigem.it
[2] Department of Neuroscience and Behavioral Sciences, University School of Medicine "Federico II",
80131 Naples – Italy
ambesi@unina.it

Keywords. CRE, algorithm, prediction, gene expression, genomics

Introduction

Biological mechanisms underlying the regulation of gene expression are not completely understood. It is known that they involve binding of transcription factors to regulatory elements on gene promoters. However, attempts to computationally predict such elements in DNA sequences of gene promoters typically yield an excess of false positives. Computational identification of CREs is currently based mainly on three different approaches: (1) identification of conserved motifs using interspecies sequence global alignments (Pennacchio 2001); (2) identification of conserved motifs in the promoters of co-regulated genes (Hughes et al 2000, Sudarsanam et al 2002, Bussemaker et al 2001, Eskin et al 2002, Bailey et al 1994, Fujibuchi et al 2001, Palin et al 2002); (3) computational detection of known experimentally identified motifs in genes' promoters for which binding factors are unknown (Kel et al 2003). The limitations of the first approach are caused by the high mutation, deletion and insertion rates in gene promoter regions (Ludwig 2002), that prevent a correct alignment of the promoter region. As experimental data is accumulating on known DNA binding elements, increasing amount of information can be used to search for similar elements in genes for which transcription factors are unknown. Our approach involves consensus pattern search of known regulatory elements in 5kb upstream of gene transcription start site against a background word distribution simulated by shuffling symbols in consensus, with the aim of minimizing false positives by using a background model of random matches of experimentally determined consensi, and integrating information from the promoters of ortholog genes.

Methods

The promoter of a human transcript specified by its RefSeq id was analyzed by downloading the sequence 4kb segment upstream plus 1kb downstream of transcription start site from *ensembl* database (www.ensembl.org), and annotations were used to identify and mask any overlapping exon from further analyses. Given a RefSeq id as input, sequences were downloaded for that transcript and for the ortholog genes stored in *ensembl* Compara 15.1 table in *ensembl* database (Clamp et al, 2003) for the mouse, rat, fugu, and zebrafish genomes.

Information on known binding factor sites were collected from TRANSFAC database version 6.4 (Heinemeyer et al, 1998), in the form of consensi which are derived from weight matrices obtained by sequence alignment of experimentally validated binding sites of transcription factors. This data is used as input for our algorithm for Binding Factor Identification (BID). A given consensus is evaluated for each of the ortholog promoters sequences by comparing the match count for the consensus to the mean match count obtained using 250 random shuffling cycles of consensus base positions. This evaluation is based on the calculation of a score *k*, according to the following formula:

$k = (mc - mcs) / sd$

in which *mc* is the number of matches found in the sequence; *mcs* is the mean match count in the shuffle cycles, and the *sd* is the standard deviation of the distribution of match counts in the shuffle cycles. Each of the ortholog promoters is evaluated independently, and a final classification of binding factor(s) characterized by the given consensus for the gene that was tested is obtained by choosing a threshold score k and a classification criterion that takes into account information from the different ortholog sequences. The final classification of a given consensus pattern was based ont the Boolean expression '*human* + *mouse* + *rat* + *fugu* + *zebrafish*'.

In order to test the performance of BID algorithm and choose the optimal parameters we randomly picked 25 test genes from TRANSFAC gene table, containing 718 human genes for which binding factors are known and used them as a test set for our algorithm. As a comparison, the performance of the web based tool Match^{TM}, available on the TRANSFAC web site (Kel et al 2003), was tested on the same set of 25 sequences, selecting the "minimize false positives" option. Match^{TM} algorithm makes use, for each weight matrix, of parameters optimized for the test genes that were also used in our analysis as a test set. Therefore a simplified version of the Match algorithm was implemented so that the test set genes could be

analyzed, for a fair comparison, with a weight matrix based algorithm with no training and no random shuffling. Specifically, in our implementation of a "simplified Match algorithm" did not use specific parameters for each matrix.

Results and Conclusion

The sensitivity-specificity plot for the BID algorithm is shown in Figure 1. This analysis was performed on the test set of 25 genes. Pairs of sensitivity and specificity values for BID were compared to the simplified implementation of match (with no training) choosing the empirically determined optimal cutoff (0.90; data not shown) and the results obtained by Match^{TM} web based tool (Figure 2). The graph suggests that BID is significantly more accurate than the simplified implementation of Match^{TM} and that its performance is very close to that of Match^{TM}.

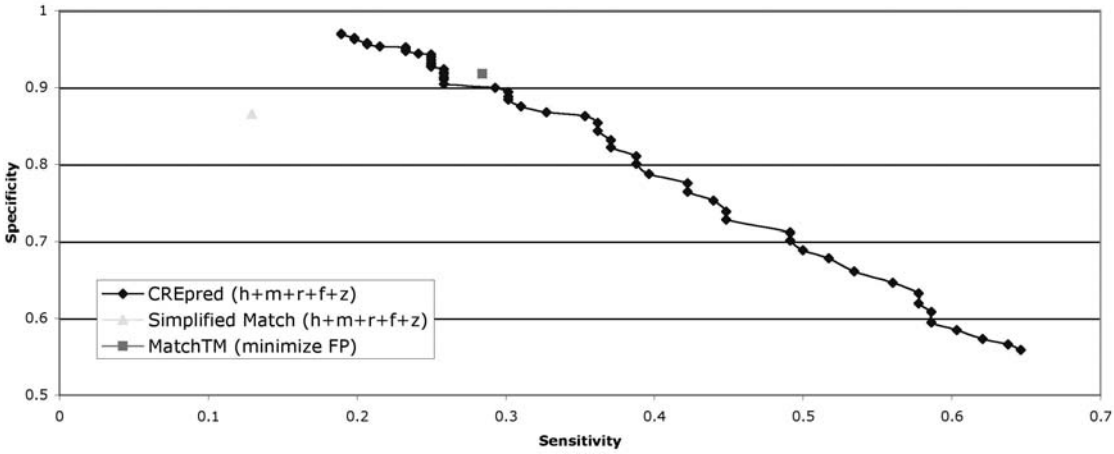

Figure 1. Performace of BID using criterion *h+m+r+f+z* is compared to that of the weight matrix search (see text) and the original MatchTM web based tool (test set genes were picked from training set used to tune MatchTM).

Our results suggest that such approach has indeed a strong potential to accurately predict CREs in human gene promoters. Integration of information from different species provided a gain in performance, suggesting that our strategy of using ortholog information is indeed fruitful. BID needs to be improved to better exploit the information from more distantly related species, such as Fugu and Zebra fish, for example taking into account the evolutionary distance, and the fact that promoters in these species have shorter sequence length. In conclusion, based on performance comparison with different algorithms of CRE prediction, our approach seems promising, and may be successfully used for the computational identification of binding factors of genes whose regulation is not known.

References

Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nature Genetics. 2001; 27: 167-171.

Fujibuchi W, Anderson JS, Landsman D (2001): PROSPECT improves cis-acting regulatory element prediction by integrating expression profile data with consensus pattern searches. Nucleic Acids Res 29:3988-96.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003): MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31:3576-9.

Palin K, Ukkonen E, Brazma A, Vilo J (2002): Correlating gene promoters and expression in gene disruption experiments. Bioinformatics 18 Suppl 2:S172-80.