Identification of human transcription factor binding sites by comparative genomics.

D. Corà[1], C. Herrmann[2], C. Dieterich[3], F. Di Cunto[4], P. Provero[5] and M. Caselle[1].

[1]Dipartimento di Fisica Teorica dell'Università di Torino and INFN ,Via P. Giuria 1 - I-10125 Torino, Italy

cora@to.infn.it, caselle@to.infn.it

[2]LGPD, Campus de Luminy Case 907 - F-13288 Marseille Cedex 9, France

herrmann@ibdm.univ-mrs.fr

[3]Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73 - D-14195 Berlin, Germany

dieteric@molgen.mpg.de

[4] Dipartimento di Genetica, Biologia e Biochimica dell'Università di Torino , Via Santena 5 bis - I-10126 Torino, Italy

ferdinando.dicunto@unito.it

[5]Fondazione per le Biotecnologie , Viale Settimio Severo 63 - I-10133 Torino, Italy

provero@to.infn.it

Keywords. Genomics, Transcription Factors.

Introduction

Understanding transcriptional regulation of gene expression is one of the greatest challenges of modern molecular biology. A central role in this mechanism is played by transcription factors (TF) which typically bind to specific, short DNA sequence motifs which are usually located in the upstream region of the regulated genes. We discuss here a simple and powerful approach for the identification of these cis-regulatory motifs based on human-mouse genomic comparison. By using the catalogue of conserved upstream sequences collected in the CORG database [1] we construct sets of genes sharing the same overrepresented motif in their upstream regions both in human and in mouse. We perform this construction for all possible words from 5 to 8 nucleotides in length and then filter the resulting sets looking for two types of evidence for coregulation: first, we analyse the Gene Ontology annotation of the genes in the set looking for statistically significant common annotation; second, we analyse the expression profiles of the genes in the set as measured by microarray experiments, looking for evidence of coexpression. The sets which pass one or both these filters are conjectured to contain a significant fraction of coregulated genes, and the upstream motifs characterizing the sets are thus good candidates to be the binding sites of the TF's involved in such regulation. In this way we find various known motifs (which we use to validate our approach) and also some new candidate binding sites.

Discussion

1. The CORG database

The CORG database [1] is a collection of conserved sequence blocks in the non-coding, upstream regions of orthologous genes from man and mouse. These blocks were obtained by searching statistically significant local suboptimal alignments of 15kb regions upstream of the translation start site. The database contains more than 10,000 pairs of orthologous genes. The alignments were obtained using the Waterman-Eggert algorithm. An important role in the following analysis is played by the fact that more than half of the genes in the database are annotated in the GO database.

2. Our analysis

We perform our analysis in four steps:

- We select only those entries of the CORG database less than 200bp long, so as to eliminate possible conserved exons. We also eliminated multiple entries so that, as a final result of this preliminary step, each nucleotide in each conserved upstream region has exactly the same statistical weight.

- Following [2,3], for each word of 5, 6, 7 and 8 nucleotides we construct the set of all genes in whose upstream region the word is overrepresented. To this end we assume as null hypothesis a random binomial distribution with a "reference probability" given by the frequency of the word in the whole database, always counting a word together with its reverse complement (i.e. we assume that the TF can act in both orientations). As a result of this second step we obtain for each word a set of genes to be examined for evidence of coregulation in the two following steps.

- For each set we consider the annotation of its genes to Gene Ontology [4] terms, the rationale being that common annotation is likely to be correlated with coregulation. Specifically, we compute (for each word and each GO term) the number of genes in the intersection between the set of genes which share the same overrepresented word in their upstream region and the set of genes annotated to the GO term in the whole human genome. We then compute the probability that an intersection as large as or larger than the one actually found occurs by mere chance, using the hypergeometric distribution.

- For each set we consider the expression profiles of the corresponding genes in a publicly available microarray dataset [5]: similar expression patterns are likely to indicate transcriptional coregulation. For each motif and each microarray, we compare the expression profile of the genes in the set with that of the entire genome, using a Kolmogorov-Smirnov test, again with a strict Bonferroni-corrected threshold on the P-value.

3. Results

The words that pass one or both coregulation tests can be easily clustered together in motifs (a similar phenomenon has been described for yeast [2,3]). Some of them turn out to be well known binding sites, but some are new. As an example of our results, one of the motifs which emerge from our analysis is GAAATTCCC which compares well with the known motif (M00054 entry of the Transfac database) GGRAAKTCCC of the NF-kB transcription factor. Besides validating our method the identification of such already known TF's is also interesting in itself since our method also gives a list of genes which are candidate to be regulated by the identified TF and a list of GO terms associated to them. Besides the known TF's we also find some new candidate binding sites which apparently were not previously identified. Needless to say, further studies and experimental tests are needed to validate these new candidates. However our method could be of great help in reducing the number of possible candidates and guide these experimental tests.

References

[1] C. Dieterich, H. Wang, K.Rateitschak, H. Luz and M. Vingron. (2003) "CORG: a database for Comparative Regulatory Genomics." Nucleic Acid Res. 31, 55-57.

[2] M. Caselle, F. Di Cunto, P. Provero. (2002) "Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes." *BMC Bioinformatics* 3:7.

[3] D. Cora', F. Di Cunto, P. Provero, L. Silengo and M. Caselle (2004) "Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs". submitted to *BMC Bioinformatics* .

[4] The Gene Ontology Consortium (2000) "Gene Ontology: tool for the unification of biology." *Nature Genetics* 25:25,29.

[5] M.L. Whitfield et al., (2002) "Identification of genes periodically expressed in the human cell cycle and their expression in tumors". *Mol Biol. Cell*, 13(6):1977-2000