Motif based classification of coexpressed genes

Pavesi G (1), Valentini G (2), Mauri G (3), Pesole G (4)

(1) Dept. of Biomolecular Science and Biotechnology, University of Milan
(2) Dept. of Computer Science, University of Milan
(3) Dept. of Computer Science, Systems and Communication, University of Milano-Bicocca
(4) Dept. of Biochemistry and Molecular Biology, University of Bari

Motivation

Understanding the complex mechanisms regulating gene expression is one of the greatest challenges for molecular biology. In particular, transcription is modulated by the interactions of transcription factors (TFs) with short DNA regions (TFBS, i.e. transcription factor binding sites) that are recognized in a sequence specific manner. The availability of genomic sequences, together with data concerning the expression of genes, has opened new opportunities in this field. In this work, we focused on the two following problems related to gene expression regulation: a) assessing whether classes of functionally related genes may be predicted using information extracted from their promoter sequences; b) the selection of motifs (TFBS) mostly related to the gene functional classes.

Methods

A quite common approach to the identification of TFBS involved in the regulation of transcription is the extraction from the promoters of a set of co-regulated (or co-expressed) genes of one or more conserved motifs, likely to represent instances of conserved TFBSs recognized by the same TF(s). Other promoters, presenting instances of the same motifs can then be predicted to be regulated in a similar fashion. The problem is that often, in the absence of experimental validation, it is very hard to assess 1) whether the conserved motifs found actually correspond to functional sites 2) which, among many candidates, are truly responsible of the regulation of the genes and, 3) if the motifs deemed to be significant (or, even those that are experimentally validated) are sufficient to explain the co-expression of the genes.

Results

We present a classification algorithm that predicts whether a gene may belong to a given set of coexpressed genes using information extracted from its promoter sequence. The classifier needs a training set composed of a set of co-expressed genes (the positive set), and a negative set, comprising genes not related to the first group. In other words, the classifier determines, from the motifs detected in the promoter of the gene, whether it is likely to be co-expressed with the genes of the positive set or not. The algorithm is composed of two steps: motif extraction and training of a classifier. In the first, motifs are extracted from each of the promoters of the genes investigated. For this task, we employed the Weeder algorithm. The main advantage is that, being Weeder an exhaustive algorithm, the result of the motif extraction phase covers every possible motif of a chosen length, and avoids the need to introduce significance criteria to select which motifs have to be used for the classification of the sequences. Moreover, Weeder is able to process each sequence separately, producing, given a motif length m, a vector composed of 4^m motif scores for each of the promoter sequences of the training set. Then, in the second step, a linear classifier is trained on the obtained score vectors. The first experiments we performed have shown very promising results. On different clusters of yeast genes with testing performed with leave-one-out cross-validation the average prediction accuracy ranged between 75% and 85%, varying according to the different clusters examined. Also, simple feature selection methods applied to the trained classifier permitted the extraction of the most significant motifs, that is, motifs that gave the greatest contribution to the correct classification of the examples. In most of the cases, the motifs matched experimentally known yeast transcription factor binding sites responsible for the regulation of the genes. We are now extending the basic idea to multi-class prediction and to other species, also evaluating the improvements deriving from the use of more sophisticated machine learning methods. While it is well known that in metazoan organisms promoter sequence alone is often not sufficient to fully

explain the patterns of expression of a gene, knowing if, and especially in which cases it can be predicted with our technique could provide valuable information and insights.

Contact email: pavesi@dico.unimi.it