A new clustering approach for learning transcriptional regulatory networks

Archetti F[1,2], Giordani I[1], Mauri G[1], Messina E[1]

[1] DISCo - University of Milano Bicocca, Milano
[2] Consorzio Milano Ricerche, Milano, Italy

Motivation

Understanding regulatory mechanisms of gene transcription is an important goal of molecular biology. To this end, an essential task is the integration of multiple heterogeneous sources of data, like measurements of RNA levels under various conditions, gene sequences, annotation, interaction data, through suitable learning methods aimed at elucidating the interactions between regulated and regulatory genes. Regulatory interactions are represented by the binding of transcription factors of regulatory genes to the promoter region of targetgenes, which contains specific motifs, recognized by transcription factors as binding sites. We can assume that co-regulated genes are likely to exhibit similar patterns of expression. Therefore common approaches to find co-regulated genes are based on gene expression clustering. This results in groups of co-expressed genes which might have similar behavior under different conditions, but that doesn't necessarily means that they are co-regulated. In order to identify sets of genes that are not only co-expressed, under different experimental conditions, but also co-regulated, a finer analysis of the upstream regions of the genes is needed. Many computational methods have been proposed with this aim, e. g. Liu et al.[7], Sinha et al. [8] try to identify regulatory elements by searching for commonly occurring motifs in the promoter regions of genes in a same cluster. Other approaches work in the opposite direction: they first extract from sequence data some predefined features of the gene, e.g. the presence or absence of some potential transcription factor binding sites, then they exploit this feature set as well as the expression data in a combined way, building models that characterize the expression profiles of groups or clusters of genes.

Methods

We propose an iterative relational clustering approach for finding groups of genes that are likely to be regulated by the same set of factors. The input is a microarray dataset for a set of genes G, where, for each gene g in G, we consider both gene expression values and DNA sequence in the upstream region of the transcription start site for g. Before submitting the microarray data to the clustering procedure, a pre-processing step, aimed at removing as much as possible the systematic noise presented in microarray data and at deleting those genes that show low variance under different conditions, is executed. The algorithm then uses the expression data in order to obtain initial clusters of co-expressed genes. Since the number k of clusters is not known in advance, the hierarchical clustering method proposed in [6] is used. This first clustering phase identifies a set of clusters, each one containing a group of co-expressed genes that are supposed to be co-regulated. In order to find the set of TFs having a high potential binding factor with promoter regions of genes in the same cluster, we use the algorithm proposed by Pavesi et al. [5], which computes the binding strength between gene and TF by means of a statistical value measuring the sequence binding specificity of known TFs. A gene TF profile is the vector of relation strength measures between the gene and the whole set of candidate TFs. This profile is used to refine the clustering model in order to optimize the extent to which the expression profile can be predicted transcriptionally. This means that a gene can be moved from a cluster to another having a closer TF profile and given these new assignments we could learn better TF models, i.e. identify the sets of TFs matching with the refined clusters, and consequently find the updated TF profiles by following the steps described above. A key advantage of our algorithm over other approaches lies in our ability to relate upstream sequence and expression data in a single framework, which allow us to refine both the cluster assignment and motifs with the same algorithm.

Results

We validated our approach on two Saccharomyces cerevisiae data sets: one [1] consisting of 173 microarrays and 6152 genes, measuring the responses to various stress conditions; the other [2] consisting of 77 microarrays and 6178 genes, measuring expression during the cell cycle. Saccharomyces cerevisiae has the advantage of being the most studied eukaryotic organi-

sm in molecular biology because it is easily modified and cultured, yet it maintains the complex regulation mechanisms of other eukaryotes. Also yeast genome is very small and it has been completely sequenced. The promoter sequence data that we considered in our experiments consist of the 500 base pairs upstream region of each gene. These sequences were retrieved from SGD (Saccaromyces Genome Database), an organized collection of genetic and molecular biological information about Saccaromyces cerevisiae [2]. We executed the algorithm with different values for the Pearson correlation threshold, and we al-ways obtained after a few iteration (say six) clusters containing almost only co-regulated genes.

References

[1] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein and P.O. Brown, Genomic expression program in the response of yeast cells to environmental changes. Mol. Biol. Cell, 11, 4241–4257, 2000

[2] P.T. Spellman, G. Sherlock, M.O. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9(12), 3273–3297, 1998

[3] Segal E., Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19:i273-282, 2003

[4] M. Clements, E. P. van Somerena, T. A. Knijnenburga and M. J.T. Reindersa, Integration of Known Transcription Factor Binding Site Information and Gene Expression Data to Advance from Co-Expression to Co-Regulation, Genomics, Proteomics & Bioinformatics Volume 5, Issue 2, Pages 86-101, 2007

[5] G. Pavesi, F. Zambelli, Prediction of over Represented Transcription Factor Binding Sites in Co-regulated Genes Using Whole Genome Matching Statistics. WILF 2007: 651-658, 2007.

[6] F. Archetti, P. Campanelli, E. Fersini, E. Messina. "A Hierarchical Document Clustering Environment Based on the Induced Bisecting k-Means". In Proceeding of the 7th International Conference on Flexible Query Answering Systems, 2006.

[7]Liu X., Brutlag D.L. and Liu J.S., Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In proceedings Pacific Symposium on Biocomputing, pages 127-38, 2001

[8] Sinha S. and Tompa M. A statistical method for finding transcription factor binding sites. In Proceedings International Conference on Intelligent Systems for Molecular Biology(ISMB),pages 344-54, 2000

Contact : mauri@disco.unimib.it