Identification of gene regulatory modules using entropy and mutual information

Di Camillo B.(1), Nair S.K.(2), Toffolo G.(1), Cobelli C.(1)

(1) Information Engineering Department, University of Padova, Padova, Italy (2)Endocrinology Division, Mayo Clinic, Rochester, Minnesota 55905, USA

Motivation

A crucial issue in microarray studies is the elucidation of how genes change expression and interact as a consequence of external/internal stimuli such as illness, drug assumption, hormone stimulation. To do so one has to reconstruct the regulatory network by describing activation/inhibition and cause-effect relationships among expression profiles. Different approaches are available in literature, but the small number of available samples with respect to the number of genes constitutes a major drawback to apply these methods to real microarray data. At present, a realistic aim is the identification of modules of gene regulation, i.e. sets of genes that are possibly regulated by the same transcription factors, or potential inhibitors or activators of a group of co-expressed genes.

Methods

A three steps method, based on Entropy and Mutual Information scores, is presented to group genes whose profiles show possible cause-effect relationships. 1) Data are discretized in three levels (up, down or not differentially expressed) with respect to a baseline, based on a noise threshold chosen so as to compromise between false positive and false negative classifications. 2) The algorithm uses Entropy and Mutual Information scores as shown in REVEAL (Liang et al., 1998) to search, for each profile x, the input genes y (the regulators) that can univocally explain the behaviour of the output gene x (the regulated gene). 3) Given the small number of samples, causative relationships can be found just by chance. In order to minimize random findings, the graph identified at the previous step is searched in order to isolate modules of highly connected genes (the high connection of a module augments the probability that at least a subset of the obtained connections is not a random finding). More in details, the algorithm searches and removes the "articulation points" and the "bridges" of the graph (an articulation point is a vertex whose removal disconnect the graph in two or more sub-graphs; a bridge is an edge whose removal disconnect the graph in two sub-graphs). The remaining sub-graphs are the output modules. The two novel aspects of the method regard data discretization step, performed in three levels on the basis of a noise threshold rather than using an arbitrary quantization of the expression range, and graph pruning step, consisting of searching for groups of highly connected genes.

Results

The method was tested through the application to synthetic data. Ability to detect modules of co-regulated genes using plain REVEAL and our algorithm were compared. Step 1 (noise-based discretization) improves both specificity and sensitivity of modules detection. Step 3 (graph pruning) improves specificity but worsen, as expected, sensitivity. Both methods performance diminishes with the number of genes and augments with the number of samples. The relative improvements of our methods with respect to plain REVEAL become more evident with the increase of genes number and the relative lack of samples. The algorithm was applied to study gene regulation in rat skeletal muscle cells stimulated with insulin. Two cultures of cell line L6 were grown: one was treated with insulin and the other used as control. Samples were collected every hour for 8 hours from both cultures, for a total of 9 time-samples per culture, and the expression level was measured using Affymetrix chips RG_U34A. Twelve modules of highly connected pseudo-genes were identified. A partial validation of the results can be obtained by comparison with a priori biological knowledge. In particular, the largest identified module contains most of the genes known to be involved in insulin regulation and insulin action. Among others this module contains Egr1 (early growth response gene), Hexokinase2, insulin receptor-related receptor (Insrr), protein phosphatase 1 catalytic subunit (likely to be related to glycogen synthase activation), plus numerous genes involved in glycolysis, electron transport chain, lipid and carbohydrate metabolism. Identified modules allow to focus the investigation on smaller sets of genes, which can be monitored in future studies and should help to better define the mechanisms involved in regulation.

Contact email: dicamill@dei.unipd.it