Detecting gene-gene interactions in microarray data using probabilistic rule sets

Chris C Holmes
Department Mathematics
Imperial College London
180 Queen's Gate
London. SW7 2AZ. UK
c.holmes@imperial.ac.uk

Submission for a Talk

Summary

A common task in microarray studies is to detect and highlight differential gene expression across different categories of tissue type. For example, tissues may be labeled as diseased/non-diseased and the object of the study may be to learn about which genes co-regulate with the diseased state and how. One approach is to treat the task within a regression analysis and consider the recorded gene expressions as predictors (covariates) and the tissue type as a class indicator. Influential predictors then relate to differentially expressed genes. Current techniques discussed in the literature typically assume a linear relationship between the gene expression levels and the tissue type. Here we describe a tailored nonlinear model that is designed to detect gene-gene interactions and allow for level dependent gene association. In particular we use probabilistic rule sets to learn the statistical dependencies in the data. That is, given a data set to model, the method infers rule-based associations between gene expressions and tissue type, producing quantitative statements of the form:

`IF` expression level of gene $j > A$ `AND` the expression level of gene $k < B$, `THEN` the log odds probability that the tissue sample belongs to class Q is increased by X:

where the parameters $\{j, >, A, k, <, B, Q, X\}$ are automatically inferred by the model using the data. In this manner the model builds up a collection of probabilistic rules that describe the various dependencies within the data. The number of genes interacting in a rule is allowed to vary, including the use of single gene main effects. Some advantage of probabilistic rule sets over other nonlinear methods are that they are (a) highly interpretable (b) allow for the inclusion of expert prior knowledge within the model and (c) automatically perform variable selection of influential predictors (genes). We use a Bayesian framework to encode prior knowledge and to account for uncertainty both in the number of rules and in the antecedents and consequences of the rules. The method is demonstrated on a number of microarray studies where it it shown to detect interesting gene-gene co-regulation and interactions. The software, written in Matlab, is available from the web site of the author.