BIOINFORMATICS

Vol. 19 no. 17 2003, pages 2210–2211
DOI: 10.1093/bioinformatics/btg298

MGraph: graphical models for microarray data analysis

Junbai Wang, Ola Myklebost and Eivind Hovig*

Tumor Biology Department, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway

Received on January 16, 2003; revised on April 3, 2003; accepted on May 21, 2003

ABSTRACT

Summary: This paper introduces a MATLAB toolbox, MGraph, which applies graphical models as a natural environment to formulate and solve problems in microarray data analysis. MGraph with its graphical interface allows the user to predict genetic regulatory networks by a graphical gaussian model (GGM), and to quantify the effects of different experimental treatment conditions on gene expression profiles by a graphical log-linear model (GLM). The power of graphical models was explored and illustrated through two example applications. First, four MAPK pathways in yeast were meaningfully reconstructed through GGM. Second, GLM was used to quantify the contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. This application may provide a valuable aid in the prediction of genetic regulatory networks, as well as in investigations of various experimental conditions that affect global gene expression profiles.

Availability: The MATLAB program MGraph is freely available at http://www.uio.no/~junbaiw/mgraph/mgraph.html for academics.

Contact: junbai.wang@radium.uio.no

INTRODUCTION

Current advances in DNA microarray technology have enabled researchers to monitor the expression levels of a large number of genes simultaneously. Such 'global' gene expression studies have widespread practical applications, ranging from yeast genetics to breast cancer research. With the development of novel technologies for genome-wide monitoring of gene expression, there is a continuous need for development of more refined statistical tools for mining of the results. This is also the case in addressing design issues of microarray experiments, where we need to interpret the causal effects of different experimental conditions having effects on global gene expression profiles. From a more ambitious perspective, we explore the possibility to infer regulatory pathways and gene interactions by the use of this methodology on expression profiles. Graphical models, which represent a combination of probability theory and graph theory, may provide a suitable tool for such studies. These are probability models for multivariate random observations whose independence structure is characterized by a graph, the (conditional) independence graph. Many statistical methods already proposed for microarray data analysis are special cases of general graphical models, i.e. mixture models, Boolean networks and Bayesian networks (Friedman *et al.*, 2000). We have developed a MATLAB toolbox, MGraph, which was implemented with two types of graphical models; analyses of continuous data summarized by a correlation matrix and discrete data summarized by a contingency table. The former was accomplished by the use of graphical gaussian models (GGM) or named covariance selection models, and the latter by graphical log-linear models (GLM).

*To whom correspondence should be addressed.

DESCRIPTION

Independence graphs

A graph G is a mathematical object that is comprised of two sets, a set of vertices, K and a set of edges, E, consisting of pairs of elements taken from K. If all edges are undirected then the graph is undirected. Let $X = (X_1, X_2, \ldots, X_k)$ be a vector of random variables, and K the corresponding set of vertices. The graph is an independence graph, or a conditional independence graph, if there is no edge between two vertices (X_i, X_j) whenever the pair of variables is independent given all the remaining variables termed 'rest'. The rest can be written as $X_i \perp X_j|\text{rest}$ (Edwards, 1995). (See figures in web supplement, http://www.uio.no/~junbaiw/mgraph)

GGMs

Given an independence graph G, and a k-dimensional continuous random vector X, a GGM is a family of normal distributions for X, constrained to satisfy the pair-wise conditional independence restrictions inherent in the independence graph. Such conditional independence constraints are equivalent to specifying zeros in the inverse variance parameter corresponding to the absence of an edge in G. In MGraph, the model selection procedure of fitting GGM is based on backward edge exclusion with a deviance different stopping rule: (1) Build the initial full graph G and the covariance matrix

cov(X) of X. (2) Compute the partial correlation coefficient matrix ρ of cov(X), and search for $\rho(i, j)$ which has the smallest absolute value among all of non-zero elements of ρ. Replace $\rho(i, j)$ as zero, then compute maximum likelihood estimates for covariance matrix cov$(X)'$. (3) Use the deviance difference, dev $= N \ln(|\text{cov}(X)|/|\text{cov}(X)'|)$ to measure the overall quality of fit for the selected model. $|\text{cov}(X)|$ is the determinate of cov(X) and N is the number of samples. The deviance dev has an asymptotic χ^2 distribution with one degree of freedom. (4) If the probability value of dev $\leq P$ (i.e. significance level $P = 0.05$), then the model selection is stopped. Otherwise, delete the edge (i, j) from graph G selected above and go back to step 2. The final selected model is an (conditional) independence graph, where vertices represent genes and edges depict associations between pairs of genes. The results of using GGM to predict four MAPK pathways in yeast (http://www.rii.com/publications/2000/s287873.htm) are displayed on our web supplement; where the associations of genes in the four pathways are meaningfully reconstructed, and results are compared by replacing missing values with zeros or by k-nearest neighbor imputation.

GLMs

In GLMs, the full range of expression data are first measured and then broken into discrete subsections (i.e. down- or up-regulated genes, e.g. by cluster analysis). These discrete subsections may be reduced to a k-way contingency table and analyzed by GLM. In the present version, model selection strategy of GLM is: (1) We have N observations (genes) of k conditions. The original continuous data are summarized to a k-way contingency table, where each discrete variable of $X = (X_1, X_2, \ldots, X_k)$ has r_i discrete subsections $(i = 1, \ldots, k)$. (2) Based on this k-way table, the initial base model is built, where the vertices correspond to k discrete variables X and the edges related to two-factor interaction terms. An edge exclusion deviance, dev$(Xb \perp Xc \mid Xa)$ is used to test the significance of pair-wise conditional independence with one missing edge; dev$(Xb \perp Xc \mid Xa) = 2 \sum n_{abc} \log[(n_{abc} n_a)/(n_{ab} n_{ac})]$, n is the marginal table of counts. The deviance has an asymptotic χ^2 distribution with $r_a(r_b - 1)(r_c - 1)$ degrees of freedom. (3) The backward stepwise edge elimination is performed to improve the base model. That is to say, initially the least significant (χ^2-test has the largest P-value) edge is removed. Then all the edges found non-significant (i.e. all $P \geq \alpha$, where α is the critical level) at step 1 are tested. The least significant edge is removed and all edges non-significant at step 2 are tested, and so on. If all P-values are significant, then no edges are removed and the procedure stops.

GLM was tested on expression data from *Drosophila melanogaster* (Wei *et al*., 2001), where one dataset contains 442 significant genes identified by the original publication and another independent dataset with 500 significant genes selected by Fisher's linear discrimination method. The results were displayed by two independence graphs (see figures in web supplement). From both graphs, the predicted associations among sex, genotype, ages and expression levels matched with the original studies, and the significance of each effect was assigned a P-value from the χ^2-test.

There are some limitations in the practical use of MGraph. First, GGM requires inverse matrix calculations that is very sensitive to the rank of the matrix, thus the number of experimental conditions should be larger than the number of genes when applying GGM to predict gene interactions. Second, the current GLM only considers the decomposable models, which have direct estimates and reduces the task of fitting all potential k-way models to fitting just 2^k two-way interactions (Edwards, 1995).

The future development of MGraph will include overcoming the present limitations, and extending undirected model selection to directed graphs or chain graphs as already available in causal inferences package developed for other types of data, i.e. TETRAD (Scheines *et al*., 1994).

ACKNOWLEDGEMENTS

This work was supported by the Norwegian Cancer Society (www.kreft.no).

REFERENCES

Edwards,D. (1995) *Introduction to Graphical Modelling*. Springer, New York.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian network to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Scheines,R., Spirtes,P., Glymour,C. and Meek,C. (1994) *TETRAD II: Tools for Discovery*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wei,J., Riley,R.M., Wolfinger,R.D., White,K.P., Passador-Gurgel,G. and Gibson,G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.*, **29**, 389–395.