Gene expression

MILVA: An interactive tool for the exploration of multidimensional microarray data

Davide D’Alimonte¹, David Lowe¹,*, Ian T. Nabney¹, Vassilis Mersinias² and Colin P. Smith²

¹Neural Computing Research Group, Aston University, Aston Triangle, Birmingham B4 7ET, UK and ²School of Biomedical and Molecular Sciences, University of Surrey, Guildford, Surrey GU2 7XH, UK

Received on July 1, 2005; revised on August 23, 2005; accepted on September 8, 2005
Advance Access publication September 13, 2005

ABSTRACT
Motivation: Clustering techniques such as k-means and hierarchical clustering are commonly used to analyze DNA microarray derived gene expression data. However, the interactions between processes underlying the cell activity suggest that the complexity of the microarray data structure may not be fully represented with discrete clustering methods.

Results: A newly developed software tool called MILVA (microarray latent visualization and analysis) is presented here to investigate microarray data without separating gene expression profiles into discrete classes. The underpinning of the MILVA software is the two-dimensional topographic representation of multidimensional microarray data. On this basis, the interactive MILVA functions allow a continuous exploration of microarray data driven by the direct supervision of the biologist in detecting activity patterns of co-regulated genes.

Availability: The MILVA software is freely available. The software and the related documentation can be downloaded from http://www.ncrg.aston.ac.uk/Projects/milva. User ‘surrey’ as username and ‘3245’ as password to login. The software is currently available for Windows platform only.

Contact: d.lowe@aston.ac.uk

1 INTRODUCTION

DNA microarray technology allows for the simultaneous measurement of the expression level of thousands of genes. Both the large quantity of data and the complex dynamics of gene expression make it difficult to identify interesting patterns of gene expression. In practice, various clustering techniques, such as k-means and hierarchical clustering, are commonly used to analyze time-course microarray experiments. However, most of the gene expression profiles are not independently related to isolated biological functions but are the result of interconnected dynamical processes undergoing the cell activity. This suggests that discrete clustering methods that split gene expression profiles into disjoint classes may misrepresent the rich structure of the data. The rationale for this work is to present a new approach that supports an exploration of the continuous structure of gene expression data on the basis of the two-dimensional (2D) topographic representation of microarray data. To address this objective we developed a software package called MILVA (microarray latent visualization and analysis). The aim of MILVA is to allow for an interactive microarray data analysis driven by the direct supervision of the biologist in detecting groups of co-regulated genes, and without the need of relying on less flexible clustering methods. MILVA is based on two recently developed topographic models: NeuroScale and generative topographic mapping (GTM).

NeuroScale (Low and Tipping, 1997; Tipping and Lowe, 1998) is a non-linear topographic model that preserves the relative similarity between the original higher dimensional data (i.e. the gene expression profiles) and their representation in a lower dimensional latent space (here represented by a plane). NeuroScale is similar in principle to the Sammon mapping (Sammon, 1969), but has the advantage of being a genuine projective model since its functional form means that it can project data not from the original training set.

The GTM (Bishop et al., 1997) approach is a fully probabilistic alternative to the self organizing map (SOM, Kohonen, 1995) and is based on the assumption that points are distributed in the proximity of a manifold embedded in the data space. A mapping from a lower dimensional latent space to the manifold allows to define a conditional probability density function in the data space. On this basis, Bayes’ theorem is then used to express the posterior distribution (i.e. lower dimensional representation) of the original data.

2 THE MILVA SOFTWARE

The MILVA software package has been developed in MATLAB on the basis of the NETLAB toolbox (Nabney, 2001). MILVA is composed of three graphical user interfaces (GUIs, Fig. 1): (1) a MAIN GUI that allows to select processing files and to define visualization options; (2) a TOPOGRAPHIC GUI for the 2D visualization of microarray data and (3) a DATA GUI for the representation of the gene expression profiles.

Points closely grouped in the TOPOGRAPHIC GUI correspond to similar patterns of gene expression. Taking advantage of this, a core feature of MILVA is to allow the exploration of gene expression patterns on the basis of their topographic representation. In fact, when the user clicks with the mouse on the TOPOGRAPHIC GUI, the set of closest points (whose size can also be specified) is highlighted and the corresponding patterns of gene expressions are visualized in the DATA GUI (e.g. as shown by the arrow linking the TOPOGRAPHIC GUI with the DATA GUI in Fig. 1). Individual gene expression patterns can also be queried (i.e. to relate each pattern of expression to the corresponding gene name and vice versa) or removed through simple mouse operations.
Notice that the joint TOPOGRAPHIC and DATA GUI visualization allows the user to identify related genes without separating microarray data into a predefined number of clusters. Additionally MILVA has the following features: (1) a set of basic filters to identify significantly expressed genes; (2) rescaling procedures; (3) gene highlighting in the TOPOGRAPHIC GUI by name; (4) gene search on the basis of a user-specified pattern of expression and (5) standard data visualization techniques (e.g. PCA) for benchmarking. For further details see the software manual available for download from the MILVA web page.

3 CONCLUSION

The strength of the proposed method is derived from the possibility of exploiting the topographic representation of microarray data for a more active exploration of the higher dimensional gene expression patterns. This and the interactive features implemented in the MILVA software allow the investigator to supervise both the data filtering process and the identification of related gene expression profiles. Subsequent analysis of microarray data can take advantage of the principled basis of the exploratory approach presented here. For instance, the interactive explorative approach that is the core of the MILVA software can be effectively exploited to investigate the dynamical similarity between gene expression profiles (D’Alimonte et al., 2005).

ACKNOWLEDGEMENTS

This work was funded under the BBSRC’s Toolkit for Functional Genomics Initiative (Grant FGT11407 to C.P.S.) and the BBSRC/EP SRC’s Exploiting Genomics Initiative (Grant 92/EGM17737 to D.L. and I.T.N.).

Conflict of Interest: none declared.

REFERENCES


Fig. 1. MILVA GUIs. The MAIN GUI on the left defines the tool settings. TOPOGRAPHIC GUI on the top right provides a 2D visualization of microarray data. Gene expression profiles are shown in the DATA GUI on the bottom right. The user can explore the data space in an informed way by navigating with the mouse in TOPOGRAPHIC GUI. Results presented in this figure are based on a microarray experiment addressed to investigate the Streptomyces coelicolor bacterium (see http://www.surrey.ac.uk/SBMS/Fgenomics/ for a description of the microarray and associated protocols).