Biochemistry 218 - BioMedical Informatics 231
Computational Molecular Biology
Final Project

Multivariate Projection Approaches for Microarray Analysis

Gang Yu

Winter 2005

Research on microarrays or gene chips presents a challenge for biologists in functional genomics: because the images generated from microarray experiments are so visually complex that manual comparisons are infeasible, computational tools are required to examine microarrays. For the past several years, there has been an explosion in the numbers of studies on microarray computational tools (Eisen et al., 1998; Burgess, 2001; Risinger et al., 2003; Wouters et al., 2003; Yeung et al., 2004; Saidi et al., 2004; Girolami and Breitling, 2004; Stoyanova et al., 2004; Tan et al., 2004; Busold et al., 2005). In general, four common themes in microarray analysis can be identified. These four themes consist of 1) detection of differential expression, 2) pattern discovery, 3) class prediction, and 4) inference of regulatory pathways and networks (Slonim, 2002).

For the pattern discovery theme, computational tools roughly fall into two major categories. One is multivariate projection methods based upon projections of high-dimensional data in a lower dimensional space and plotting both genes and samples in this lower dimensional space using the biplot (Chapman et al., 2002). This projection into a subspace of low dimensionality can account for the main variance in the data. The other is cluster analysis methods.

This paper discusses approaches in the first category, multivariate projection methods; however, tools in the second category may be mentioned for comparisons. The first part of the paper provides a brief overview of the multivariate methods. Then, algorithms of several major multivariate approaches are presented in the second part. The following part summarizes advantages and drawbacks of multivariate methods with a comparison with cluster tools. The final part is a conclusion.

I. A Brief Overview of Multivariate Projection Approaches

The multivariate projection methods include principal component analysis (PCA), correspondence factor analysis (CFA), spectral map analysis (SMA), partial least squares (PLS) method, and some other variants. Initially, all of these methods were developed in either statistics or other academic areas, but recently used in microarray analysis. Multivariate projection methods help to reduce the complexity (dimensions) of highly dimensional data (n genes versus p samples) and provide means to identify gene patterns or subjects in the data. Projected data are typically displayed in a biplot (genes and samples) in a new space.

PCA is the oldest and best known of the multivariate projection techniques. Historically, PCA dates back to Pearson (1901) and Hotelling (1933). This approach tries to identify components that explain the variance in the data. The central idea of PCA is to reduce the dimensionality or complexity of a data set, while retaining as much as possible of the variation present in the data. The dimension reduction technique is accomplished by introducing a new set of variables "principal components" that are linear combinations of

the original variables and uncorrelated to each other. In other words, PCA reproduces the total variance among a large number of variables using a much smaller number of unobservable variables or dimensions called latent factors. Principal components can be determined with different methods such as singular value decomposition (SVD) or some other algorithms. For the just past three years, numerous papers, for example, Peterson, (2003), Barra (2004), Saidi et al., 2004, Tham et al. (2003), Girolami and Breitling (2004), and Hubert and Engelen (2004) have applied this method in microarray studies.

In early 1970s, J. P Benz´ecri developed CFA method for contingency tables and in a sense decomposed the χ^2 statistic. Therefore, distances between objects in CFA have a χ^2 distribution. The method has been widely employed to multivariate data analysis in sociology, environmental science, and marketing research. Kishino and Waddell (2000), Fellenberg et al. (2001), Peterson (2002), Tham et al. (2003), Perelman et al. (2003), Wouters et al. (2003), Tan et al. (2004), and Busold et al. (2005) introduced the method to the investigations of microarray data by displaying the associations between genes and experiments. Since CFA was primarily designed for analyzing contingency tables, it can reveal the association both between and within all the variables (genes and experiments) simultaneously.

Like CFA, SMA was originally developed in 1970s. This method was developed not for biological research either, but for the display of activity spectra of chemical compounds (Lewi, 1976). In the past, SMA has been successfully applied to a wide variety of problems, ranging from pharmacology (Lewi, 1976), virology (Andries et al., 1990), to management and marketing research (Faes and Lewi, 1987). Thielemans et al. (1988) have compared SMA with PCA and CFA, using a relatively small data set from the field of epidemiology. Recently, Wouters et al. (2003) and Peeters et al. (2004) applied this multivariate projection method to microarray analysis. They all argued that SMA would be a promising new tool for microarray data analysis.

PLS is another well-known dimension reduction technique. Wold (1975a, 1975b) developed the PLS approach initially used for modeling information-scarce situations in social science but recently employed in biochemistry. The method relates the data matrix X to a y-response that can be either a single y or multiple Y, i.e., generating a model that predicts y or Y from X. In the computer literature jargon, PLS is known as a supervised method in that it uses both the independent and the dependent variables, whereas PCA is an un-supervised method that considers only independent variables. Datta (2001), Nguyen and Rocke (2002), Park et al. (2002), Johansson et al. (2003), Pérez-Enciso and Tenenhaus (2003), Man et al. (2004), Tan et al. (2004), and Nguyen (2005) have applied this statistical method to microarray data analysis.

II. Algorithms

Due to space limit of this paper, exhaustive explanations of these methods will not be presented; however, a review of the basic elements and major structures of their algorithms is necessary to understand their capabilities in microarray analysis. Concise presentations of the algorithms will be given below.