Management and Statistical Analysis of Microarrays Data

C. Romualdi, B. Celegato, S. Campanaro, N. Cannata, S. Toppo, G. Valle, G. Lanfranchi

CRIBI, Università di Padova

One of the main current activities of the CRIBI Genomic Group is the application of cDNA microarray technology for studying the pattern of gene expression in different physiological and pathological condition of skeletal muscle (http://muscle.cribi.unipd.it/microarrays). At the bioinformatic level this work can be divided into two sections: data management and statistical analysis.

Data management has been approached by means of a relational database with SQL structure where samples, spots and raw/normalized data are linked to each other through a series of SQL tables. Thus, it is possible to retrieve all the information concerning any specific clone to the relevant experiment results, and vice versa. This ordered system allows sistematic queries about gene expression pattern. In particular a dedicated web interface (http://muscle.cribi.unipd.it/expression) has been developed to display the results from some microarray experiments. The simple two-fold method for the identification of the differentially expressed gene is implemented, and a list of over/under expressed genes are reported with their expression levels and MUSCLE-TRAIT entries.

Over the last two years a series of sophisticated computational tools were developed for the analysis of different aspects of gene profiling (discriminant analysis, neural networks and support vector machine). In consideration of the on-going microarray projects in our lab, there was the need to compare supervised statistical techniques on the basis of their effectiveness in correctly classifying different physiological and pathological conditions.

The purpose of our work was the assessment of the performance of five selected discriminant analysis techniques using either a simulation approach or experimental datasets. The error rate found by cross validation has been used to evaluate the statistical methodologies. A simulation approach is significantly important for this comparison, because it allows evaluating the statistical methodologies in simplified, predefined and piloted situations, so that the interpretation of the results can be greatly facilitated. In particular we can control the huge source of variation among and between pathological conditions and we can evaluate methodologies under such controlled experimental conditions. Through simulated matrices we infer the minimum number of affected patients (30) and genes (20) required for a good classification. The results obtained with simulated and real dataset agree in identifying the linear discriminant analysis as the best methodology in term of misclassification rate.