An Approach to Multifactorial Microarray Data Analysis

ID - 219

Maratea Antonio[1], Michele Ceccarelli[1], Pierre Baldi[2]
[1]Research Center On Software Technology, University of Sannio, via Traiano 11, Benevento
[2]Institute for Genomics and Bioinformatics, University of California, Irvine, USA

Motivation

Multifactorial microarray analysis is a challenging task in many ways.
When sampling conditions increase in size with the aim of taking into account more possibly related variables, isolated effects of each variable are progressively hidden by the added noise and the dependence relations introduced. Microarray experiments are usually arranged in a matrix D of size MxN where M is the number of genes and is in the order of tens of thousands and N is the number of experimental conditions and is in the order of tens. In one-factor experiments, N are levels of one variable (like hours if the factor is time); in multifactorial experiments, the N conditions are joint-levels of two or more possibly related variables (for example each condition is a measure at a given time and concentration, if the factors are time and dose). In the latter case, the level of expression of each gene can be represented as a surface in a suitable k-dimensional space, where (k-1) are the experimental variables and the k-th dimension is the expression's value (for example the x and y axis are time and doses and the z axis is the level of expression).
What we may be interested on is to find over-represented patterns of co-variation of genes among various conditions, that is 'surface modes', rather than isolating single genes significantly affected by each variable under study.

Methods

The approach we took here is similar to a data reduction technique, but as we are interested in reducing the data row-wise, that is in removing uninteresting genes' profiles instead of uninteresting experimental conditions, we have tens of points in a space with tens of thousands of dimensions and this fact prevents us from the use of standard techniques like PCA or ICA. A first step of the method is data coding: given the matrix D of size MxN, each value is substituted with the integer that represents its quartile of belonging, plus a fifth class for extreme values coded as 5 (a finer decomposition it's of course possible using other percentiles). Starting from the coded matrix C of size MxN with Cij belonging to (1,...,5), for each gene Gi, a single number that measures how much its pattern of expression is present in the whole dataset is computed and saved in component i of a vector S. The computation of the overall presence of a given gene pattern in the whole dataset is done through a suitable measure of similarity mutuated from sequence analysis. Si shows how much the profile of gene Gi is represented in all the data. In the next step the gene corresponding to the maximum value of S is choosen, let's call it G-MAX, and the similarity of this gene's profile with respect to each other gene individually is computed and saved in component i of vector Q. Qi is the measure of similarity of gene G-MAX with respect to gene Gi. Data are hence ranked according to the values of Q and all the genes that are under the a*max(Q) threshold, where $0<a<1$, are considered to be the first 'mode'. The second 'mode' is found removing the genes beloging to the first 'mode' from the dataset and repeating the whole process, apart coding. So on for the other 'mode'.

Results

Applying this method to a large toxicology experiment we obtained promising results. The two variables investigated toghether were time (4 levels) and doses (5 levels) and we had for one chemical a total of 24 combined measurements of them, considering also controls (dose 0 at each time, 4 levels). First we performed differential expression analysis and then we applied the method both with respect to the reduced list of differentially expressed genes than with respect to the complete gene list. In both cases the method is able to recognize meaningful patterns corresponding to responses for example to high doses, high early responses etc. Comparing the results of the analysis on the differentially expressed reduced gene list and on the complete gene list we see that while in the former case the meaningful 'modes' are the first ones, in the latter case the meaningful 'modes' are the last ones. This fact matches the expectation that the majority of genes do not react and confirms the effectiveness of both differential expression analysis and the proposed 'modes' method.
This method has three parameters: one regarding quantization scale; another regarding the number of top ranked points to keep in each meaningful 'mode' and another regarding how many meaningful 'modes' to keep. In the studied case we found that reasonable values produced best results and that the interpretation of the resulting modes was quite straightforward. A deeper study of parameters'effect in more critical cases is to be performed.

Email: amaratea@unisannio.it