Identifying statistically significant patterns of expression via Bayesian Infinite Mixture Models

Mario Medvedovic
University of Cincinnati Medical Center
3223 Eden Av. ML 56 Cincinnati, OH 45267-0056
USA
513-558-8564
513-5584838
medvedm@email.uc.edu
CAMDA00 Dataset 1: Yeast

Mario Medvedovic

Identifying statistically significant patterns of expression via Bayesian Infinite Mixture Models
Mario Medvedovic, University of Cincinnati Medical Center

Generally, the expression data obtained from DNA arrays incorporate several sources of variability present in the process of obtaining fluorescence intensity measurements (15). When looking at expression profiles of thousands of genes at once, certain groups of genes will exhibit some level of similarities purely due to random chance. Such spurious results are almost inevitable unless a proper statistical model is applied to assess the statistical significance of the observed patterns. Assessing statistical significance of observed expression patterns means determining the level of similarity that is unlikely to be the result of random fluctuations in observed data. The common denominator of all clustering procedures currently used in the analysis of microarray data is their inability to establish statistical significance of observed clusters.

We have developed a clustering procedure based on the Multivariate Gaussian Bayesian Infinite Mixture model. The basic idea in this approach to identifying distinct patterns of expressions is to define a probabilistic model for the clusterings of observed gene expression profiles. Using Gibbs sampler, the posterior distribution of clusterings is simulated by generating a sequence of clusterings with distribution that approximates the posterior distribution of clusterings under our probabilistic models.

Suppose that T gene expression profiles were observed across M experimental conditions. Let xki denote the expression level of the ith gene for the kth experimental condition and xi=(x1i, x2i,…, xMi) is the set of all expression levels for the ith gene. That is, xi denotes the complete expression profile for the ith gene. If ci is the classification variable indicating the cluster to which the ith expression profile belongs (ci=k means that the ith expression profile belongs to the kth cluster), then a "clustering" is defined by a set of classification variables for all expression profiles c=(c1, c2,…, cT). The values of classification variables are meaningful only to the extend that all observed expression profiles having the same value for their classification variable form a cluster. Let Q denote the number of clusters defined by In the Bayesian Infinite Mixture model the number of clusters (Q) as well as the classification variables (c1,…, cT) are considered to be random variables. Instead of finding a single optimal clustering, the posterior distribution of classification variables given data is estimated. Clustering based on posterior distribution of all possible clusterings is averaged over models with all possible number of components. In addition to automatically detecting significant clusters without the need to specify number of clusters prior to the analysis, this model allows us to impute data for genes that for which differential expression was not observed in all experiments. The cell-cycle database will be analyzed first by using only profiles with observed expressions in all experiments, and second by using all profiles and by imputing missing data.

Keywords

Bayesian, Mixture Model, Dirichlet Process, Microarray

Tools

website

http://homepages.uc.edu/~medvedm/