

Statistical inference for simultaneous clustering of gene expression data

Katherine S. Pollard

Current methods for analysis of gene expression data are mostly based on clustering and classification of either genes or samples. I offer support for the idea that more complex patterns can be identified in the data if genes and samples are considered simultaneously. Visualization methods help to illustrate such patterns. I formalize the approach and propose a statistical framework for simultaneous clustering. This framework allows one to assess classical properties of clustering methods, such as consistency, and to formally study statistical inference regarding the clustering parameter. I present results of simulations designed to assess the asymptotic validity of different bootstrap methods for estimating the distribution of clustering parameters.

For both hierarchical and partitioning clustering algorithms, selecting the number of significant clusters is an important problem and many methods have been proposed. Existing methods for selecting the number of clusters tend to find only the global patterns in the data (e.g.: the over and under expressed genes). A better method is needed in the gene expression context, where small, biologically meaningful clusters can be difficult to identify. I propose to choose the number of clusters as the minimizer of cluster heterogeneity and suggest several criteria for measuring heterogeneity in the clustering context. The power of this method compared to existing methods is demonstrated on simulated and experimental microarray data.