EXCAVATOR: A new software for gene expression data analysis

Dong Xu, Victor Olman, Li Wang, and Ying Xu
Protein Informatics Group, Life Sciences Division
Oak Ridge National Laboratory
xud@ornl.gov

Abstract

As one of the most popular tools for genome analysis, microarrays have made it possible to simultaneously monitor the expression levels of thousands of genes under different experimental conditions. Massive gene expression data have been generated using microarrays. Clustering correlated gene expression data is useful for studying the functional relationship between the genes in a biological process. We have developed a new software, EXCAVATOR, for clustering gene expression profiles based on our new framework for representing gene expression data as a minimum spanning tree. EXCAVATOR uses a number of rigorous and efficient clustering algorithms, including one with guaranteed global optimality. It can cluster thousands of genes efficiently on a PC or a workstation. It has a number of additional unique features, including capabilities for (i) automatic selection of the most "natural" number of clusters, (ii) data-constrained clustering, (iii) identification of genes with similar expression profiles to a set of specified seed genes, (iv) cluster identification from background noise, and (v) comparing different clustering results of the same data set. EXCAVATOR can be run from a Unix/Linux shell or from a Java interface, which a user can visualize clustering results in colored figures and 2D plots. It provides a wide range of options for data formats, distance measures, and clustering algorithms etc. To demonstrate the effectiveness of EXCAVATOR, we have tested it on several experimental data sets, including expression data from yeast Saccharomyces cerevisiae, human expression data related to Leukemia and in response of fibroblasts to serum, and Arabidopsis expression data in response to chitin elicitation.