Chapter LXI
C-MICRA:
A Tool for Clustering Microarray Data

Emmanuel Udoh
Indiana University–Purdue University, USA

Salim Bhuiyan
Indiana University–Purdue University, USA

INTRODUCTION

In the field of bioinformatics, small to large data sets of genes, proteins, and genomes are analyzed for biological significance. A technology that has been in the forefront of generating large amounts of gene data is the microarray or hybridization technique. It has been instrumental in the success of the human genome project and paved the way for a new era of genetic screening, testing, and diagnostics (Scheena, 2003). The microarray data set can be made of thousands of rows and columns. It often contains missing values, exhibits high-dimensional attributes, and is generally too large for manual management or examination (Tseng & Kao, 2005; Turner, Bailey, Krzanowski, & Hemingway, 2005). Database technology is necessary for the extraction, sorting, and analyzing of microarray data sets.

A plethora of techniques from machine learning, pattern recognition, statistics, and databases has been deployed to address the issue of knowledge discovery from large microarray data sets. Techniques commonly employed in this endeavor are data transformation, scatter plots, principle component analysis, expression maps, pathway analysis, workflow management, support vector machines, artificial neural networks, and cluster analysis (Baxevanis & Ouellette, 2005; Rogers, Girolami, Campbell, & Breitling, 2005). These techniques can be categorized into two groups: supervised or unsupervised methods. The supervised methods, such as support vector machines and linear discriminant analysis, require two sets of measurements. The first set represents the expression measurements from the microarray gene chips run on a set of samples, while the second is the data characterizing the samples under investigation. The goal of this method is to train the model for predictive purposes. Supervised methods tend to determine genes that fit a predetermined pattern. For a broad overview of supervised methods, the reader is referred to Scheena (2003). The second category of approaches is the unsupervised method, and it characterizes components of a data set without a prior input or knowledge

of a training signal. Clustering is an example of an unsupervised method. While these techniques are useful in microarray analysis, the focus of this article will be on clustering.

Clustering is an important unsupervised method in the exploration of the expression patterns in microarray analysis. As a tool of discovery, clustering classifies similar objects into different groups or nonoverlapping clusters so that the data in each group share commonality, often proximity according to some defined distance measure. Clustering algorithms can be partitional or hierarchical. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once (Bolshakova & Azuaje, 2006). These algorithms can be used to determine what group a particular genetic sample belongs to and the tendency for certain clusters to be associated with certain characteristics. The most widely used clustering techniques are hierarchical, *k*-means, and self-organizing maps. In the literature, other terms used interchangeably for clustering are cluster analysis, automatic classification, numerical taxonomy, botryology, and typological analysis (Fung, Ye, & Zhang, 2003; Glenisson, Mathys, & De Moor, 2003).

BACKGROUND

Numerous data mining studies based on partitioning groups in multidimensional microarray data sets reveal the significance of clustering in biological information extraction (Au, Chan, Wong, & Wang, 2005; Bolshakova & Azuaje, 2006; Lacroix, 2002; Piatetsky-Shapiro & Tamayo, 2003). A broad overview of biostatistical clustering approaches in microarray analysis is given by Scheena (2003). There are large clustering algorithms in the literature, but they are relatively equivalent in performance.

Clustering techniques partition data that are not a priori known to contain any identified subsets and can determine intrinsic grouping in a set of microarray data using distance or conceptual measures. To determine membership in a cluster, clustering algorithms evaluate distance between a point and the cluster centroid. The output is basically a statistical description of the cluster centroid with the number of components in each cluster. It is a data reduction method in that the observations in a group can be viewed as a mean of the observations in that subset. However, domain knowledge is useful to formulate appropriate measures in a clustering algorithm, which may be exclusive, overlapping, hierarchical, or probabilistic (Hanczar, Courtine, Benis, Hennegar, Clement, and Zucker, 2003; Parsons, Haque, & Liu, 2004).

Programs such as Cluster (Eisen, Spellman, Brown, & Botstein, 1998) and Hierarchical Clustering Explorer (HCE; Seo & Schneiderman, 2002) provide useful insights on summarized representations of groups in microarray data. Eisen et al. implemented in Cluster a hierarchical clustering (HC) algorithm based on the average linkage method of Sokal and Michener, which was developed for clustering correlation matrices. Although Cluster implements hierarchical clustering, *k*-means, and self-organizing maps efficiently, it is generally known to be weak in visualization. To strengthen Cluster, an interactive graphical program TreeView was developed to visualize the results of Cluster. It allows tree and image-based browsing of hierarchical clusters. Another program with interactive capability is the HCE. This program implements only hierarchical clustering, but it has more functionalities than Cluster and TreeView. For instance, HCE implements several techniques involving entire data sets, dynamic query controls, coordinated displays, scatter plots for relevance ordering, gene ontology browsing, and profile search with temporal patterns. Based on the method developed by Yeung, Haynor, and Ruzzo (2001), HCE can be described to have high predictive power. The C-MICRA (Clustering Microarray) program developed by the authors has several features available in HCE except scatter plots and gene ontology browsing, and it manages scrolling and cluster comparison better than HCE.