Comparison of commonly used algorithms for clustering gene expression data

Catherine Beauheim[1], Armin O. Schmitt[2] and Subha Srinivasan[1]
Berlex Biosciences[1], 15049 San Pablo Avenue, Richmond, CA 94804 USA
metaGen[2], Gesellschaft für Genomforschung, Ihnestraße 63,
14195 Berlin, Germany
catherine_beauheim@berlex.com, armin.schmitt@metagen.de,
subha_srinivasan@berlex.com

Extraction of biologically meaningful information from a large number of gene expression experiments using microarray technology remains a challenge. Clustering is a commonly used data mining tool to extract relevant information from large sets of data in many disciplines. However, it is not clear *a priori*, which of the many described algorithms in literature are appropriate for expression data. Commonly used clustering algorithms are computationally intensive when used with large data sets. Since of the thousands of genes on the chip only a subset carries information relevant to the biological state under study, it may be possible to eliminate redundancy from the data set prior to clustering. Here we explore the use of similarity-exhaustion based grouping (SimBG) to reduce the data set.

SimBG is commonly used to cluster large DNA sequence databases because of speed. More recently, SimBG is implemented in several commercial software programs for clustering large expression data obtained from both Affymetrix and Synteni technologies. We compare the results from the SimBG method implemented in GepMine, a proprietary software system developed for analyzing gene expression data, with other commonly used clustering methods such as Fitch, Bayesian and SOM. For comparison we use an expression data set obtained from the literature where expression of 112 developmental rat genes have been monitored at 9 time points during CNS development (http://rsb.info.nih.gov/ mol-physiol/PNAS/GEMtable.html.). The results of any two given clustering methods are compared by means of contingency table analysis and measures of association such as Cramer’s V, Tchouprov coefficient, and equivalence match coefficient. We observe that clearly distinct clusters can be extracted quickly using SimBG, suggesting that this method can confidently be used on large data sets to extract obvious clusters and to eliminate redundancy. If needed, such a reduced set can further be analyzed using more rigorous and sensitive clustering methods.