Abstract

Clustering is one of the fundamental processes of analyzing gene expression data, basically by comparing gene expression profiles or sample expression profiles. Comparing expression profiles requires a measure apart from the actual clustering algorithm to quantify how similar or dissimilar the objects under consideration are. Various clustering algorithms have been used to analyze gene expression data. Some of these algorithms reported the incorporation of similarity measures like Euclidean Distance, Pearson Correlation and mutual information for their performance. This work considered different reported clustering algorithms for gene expression data analyses and the importance of different similarity measures for optimizing these clustering algorithms. To this end, no clustering technique in all the works investigated has been applied directly on gene expression data. It is observed that the output (distance matrix) of similarity or dissimilarity measures plays the role of input to clustering techniques, and those that did not use any of the popular proximity measures applied one or two approaches such as Constrained Coherency (CoCo), Silhouette coefficient measurement, and normalization and discretization, to refine gene expression data for improved cluster quality by speeding up the learning phase,
reduction of computational space and handling of noise effectively.

References


**Index Terms**

Computer Science, Artificial Intelligence

**Keywords**

Clustering Algorithms, Proximity Measure, Gene Expression Data, Distance Matrix, Microarray.