Details of the Clustering Algorithms Supplement to the paper "Validating Clustering in Gene Expression Data" (to appear in Bioinformatics)

Ka Yee Yeung, David R. Haynor, Walter L. Ruzzo

October 16, 2000

We implemented three partitional clustering algorithms: the *Cluster Affinity Search Technique* (CAST) [Ben-Dor et al. 1999], an *iterative* partition algorithm and the *k-means* algorithm [Jain and Dubes 1988]. Three hierarchical clustering algorithms were also implemented: single-link, average-link and complete-link. For comparison, we also implemented random clustering. K-means, iterative and random are randomized algorithms; the others are deterministic.

1 Partitional Algorithms

1.1 CAST

The Cluster Affinity Search Technique (CAST) is an algorithm proposed by [Ben-Dor et al. 1999] to cluster gene expression data. The input to the algorithm includes the pairwise similarities of the genes, and a cutoff parameter t (which is a real number between 0 and 1). The clusters are constructed one at a time. The current cluster under construction is called C_{open}. The *affinity* of a gene g, $a(g)$, is defined to be the sum of similarity values between g and all the genes in C_{open}. A gene g is said to have high affinity if $a(g) \geq t|C_{open}|$. Otherwise, g is said to have low affinity. Note that the affinity of a gene depends on the genes that are already in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open}. C_{open} is *closed* when no more genes can be added to or removed from it. Once a cluster is closed, it is not considered any more by the algorithm. The algorithm iterates until all the genes have been assigned to clusters and the current C_{open} is closed.

When a new cluster C_{open} is started, the initial affinity of all genes are 0 since C_{open} is empty. One additional heuristic that the authors [Ben-Dor et al. 1999] implemented in their software BIOCLUST is to choose a gene with the maximum number of neighbors to start a new cluster. Another heuristic is that after the CAST algorithm converges, there is an additional iterative step, in which all clusters are considered at the same time, and genes are moved to the cluster with the highest average similarity.

1.2 Iterative Partition Algorithm[1]

The input to the iterative partition algorithm consists of a similarity matrix S, and a parameter α. Varying the parameter α produces clustering results with different numbers of clusters. The total similarity of a gene g to a cluster C, $Sim(g, C)$, is defined as the sum of the pairwise similarities from g to each gene in C, *i.e.*, $Sim(g, C) = \sum_{x \in C} S(g, x)$, where $S(g, x)$ is the pairwise similarity of gene g and gene x. The *excess similarity* from a gene g to a cluster C is defined as the excess of the total similarity from g to C over α multiplied by the size of cluster C, *i.e.*, $Sim(g, C) - \alpha * |C|$.

[1] We would like to thank Richard M. Karp at University of California at Berkeley for suggesting this clustering algorithm.

Initially, each gene is in its own cluster. A random order is selected for the genes in the iterative step. In each iteration, for each gene g, the excess similarity from gene g to each existing cluster is computed. If C_{max} is the cluster with the maximum excess similarity to gene g and gene g is not currently in cluster C_{max}, gene g is removed from the cluster it is in, and is inserted in cluster C_{max}. This process is repeated until no genes are moved between clusters.

This algorithm is similar in some ways to the iterative step in CAST. One difference is that in CAST, there is only one cluster open at a time, while all clusters are open at the same time in the iterative algorithm.

1.3 K-means

The number of clusters, k, is an input to the k-means clustering algorithm. Clusters are described by *centroids*, which are cluster centers, in the algorithm. In our implementation of k-means [Jain and Dubes 1988], the initial centroids consist of k randomly chosen genes. Each gene is assigned to the centroid (and hence cluster) with the closest Euclidean distance. New centroids of the k clusters are computed after all genes are assigned. The steps of assigning genes to centroids and computing new centroids are repeated until no genes are moved between clusters.

2 Hierarchical Algorithms

Agglomerative hierarchical algorithms build clusters bottom up. Initially, each gene is in its own cluster. In each step, the two clusters with the greatest cluster similarity are merged. This process is repeated until the desired number, k, of clusters is produced. Different cluster similarity criteria yield different clustering algorithms. The algorithms also naturally define a dendrogram (tree) relating the clusters and/or subclusters. Refer to [Jain and Dubes 1988] for detailed discussions on hierarchical algorithms.

2.1 Single-Link

In hierarchical single-link clustering algorithm, the cluster similarity criterion is the maximum similarity between a pair of genes, one from each of the two clusters to be merged.

2.2 Average-Link

In average-link, the cluster similarity criterion is the average pairwise similarity between genes in the two clusters.

2.3 Complete-Link

In complete-link, the cluster similarity criterion is the minimum similarity between a pair of genes, one from each of the two clusters.

3 Random Clustering

As one benchmark for evaluating the performance of a clustering algorithm, we can compare it to random clustering. A random clustering with k clusters is obtained by placing k randomly selected genes into separate bins, then placing the remaining genes into the same bins uniformly at random. An algorithm whose figure of merit is little better than that of a random clustering is probably producing poor clusters.