Clustering microarray data

Clustering gene expression data: understand gene function, gene regulation, cellular functions, subtypes of cells, etc.
- coexpressed genes have similar cellular functions

Data to be clustered: obtained from multiple experiments conducted under varying conditions
For instance: a time series of a cellular process (each experiment is a snapshot in time of gene expression levels)

Expression matrix: $M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mm} \end{bmatrix}$
- $W_{ij}$ indicates a gene
- $w_{ij}$ indicates a sample
- $m_1$ = the # of genes
- $m_2$ = the # of samples

Rows of $M$: expression patterns of genes
Columns of $M$: expression profiles of samples

Def. Clustering is the process of grouping data objects into a set of disjoint classes (clusters), so that objects within a class are highly similar while those in different sets are dissimilar

Clustering - unsupervised classification (no predefined classes)

Categories of clustering:
1. gene-based clustering (genes are objects, samples are features)
2. sample-based clustering (samples are objects, genes are features)
3. subspace clustering (genes and samples treated symmetrically)
   - cluster a subset of genes across a subset of samples

The three categories face very different challenges
Measure of proximity

Quantify the distance between two objects \( O_i, O_j \); define vectors \( O_i = \{ o_{ik}, 1 \leq k \leq p \} \), \( p \): the number of features

Euclidean distance:

\[
E(O_i, O_j) = \sqrt{\sum_{d=1}^{p} (o_{id} - o_{jd})^2}
\]

For various computational reasons, \( O_i \)'s are first standardized with zero mean and unit variance:

\[
O'_{id} = \frac{O_{id} - \mu_{O_i}}{\sigma_{O_i}}; \quad \mu_{O_i} = \frac{1}{p} \sum_{d=1}^{p} O_{id} \quad \sigma_{O_i} = \frac{1}{p-1} \sum_{d=1}^{p} (O_{id} - \mu_{O_i})^2
\]

An alternate measure - Pearson's correlation coefficient:

\[
P(O_i, O_j) = \frac{\sum_{d=1}^{p} (O_{id} - \mu_{O_i})(O_{jd} - \mu_{O_j})}{\sqrt{\sum_{d=1}^{p} (O_{id} - \mu_{O_i})^2 \sum_{d=1}^{p} (O_{jd} - \mu_{O_j})^2}}
\]

where \( \mu_{O_i} = \frac{1}{p} \sum_{d=1}^{p} O_{id} \), \( \mu_{O_j} = \frac{1}{p} \sum_{d=1}^{p} O_{jd} \).

Here, each object is treated as a random variable with \( p \) observations, and similarity is computed in a statistical sense.

Challenges of gene clustering

Features of a good clustering algorithm:

1. Should depend as little as possible on prior knowledge
2. Should be robust with respect to noise
3. Should allow for high intersection between clusters
4. Preferably, can discover relationship between genes in the same cluster
Categories of clustering:

1. gene-based clustering (genes are objects, samples are features)
2. sample-based clustering (samples are objects, genes are features)
3. subspace clustering (genes and samples treated symmetrically)

Today: clustering algorithms

1. partitional clustering (predetermined # of clusters)
   - K-means, fuzzy C-means clustering
2. hierarchical clustering
   - agglomerative (bottom-up), divisive (top-down)

K-means

A typical partition-based clustering method: given a prespecified number K, partition the data into K disjoint subsets to min.

$$E = \sum_{i=1}^{K} \sum_{0 \in C_i} |0 - \mu_i|^2$$, where

0: a data object in a cluster Ci
\(\mu_i\): the centroid (mean of objects) of Ci

objective: min. the sum of the squared dist. of objects from their cluster centers

The algorithm:

0. Initialize: randomly assign data to K sets
1. Compute centroids for each set
2. Re-assign each point to a cluster with a centroid closest to that point
3. Repeat 0, 2, until convergence

The time complexity: \(O(lKKn)\)

l - the number of iterations, K - the number of clusters
n - the number of objects (genes)
Downsides:

1. $K$ is typically unknown, so need to try out a few different
   (may end up being computationally very complex)
2. Also, may be sensitive to noise (since it forces every
   object/gene into a cluster)

### Fuzzy C-means clustering

Allow each object to belong to 2 or more clusters:
- A point close to the center of a cluster "belongs more" than the
  one close to the edge (as in fuzzy logic)

Denote: $u_j(x_i)$ - the degree of membership of $x_i$ to the cluster $j$,

$$\sum_{j=1}^{C} u_j(x_i) = 1, \quad C - \text{the total # of clusters}$$

The centroid of the $j^{th}$ cluster is defined as:

$$c_j = \frac{\sum_{i=1}^{m} u_j(x_i) \cdot x_i}{\sum_{i=1}^{m} u_j(x_i)}, \quad 1 \leq m, \quad u_j(x_i) = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_k\|}{\|x_i - c_k\|} \right)^{2/(\ell-1)}}$$

Objective function to optimize:

$$J_m = \sum_{i=1}^{m} \sum_{j=1}^{C} u_j(x_i) \cdot \|x_i - c_j\|^2$$

The algorithm:

1. Initialize $U^{(1)} = \{u_j(x_i)\}$

2. At $k^{th}$ step: use $U^{(k)}$ to compute $c_j = \frac{\sum_{i=1}^{m} u_j(x_i) \cdot x_i}{\sum_{i=1}^{m} u_j(x_i)}$

3. Update $U^{(k)}$:
   
   $$u_j(x_i) = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_k\|}{\|x_i - c_k\|} \right)^{2/(\ell-1)}}$$

4. If $\|U^{(k)} - U^{(k-1)}\| < \varepsilon$, stop; otherwise, go to 2.
Determining the number of clusters

A heuristic choice: \( K \approx \sqrt{m} \)

Other partition-based techniques

Self-organizing map (SOM) - neural network based

Hierarchical clustering algorithms

Partition-based clustering (K-means, fuzzy C-means): directly decompose the data into a set of disjoint clusters

Hierarchical clustering: generates a series of nested clusters, graphically represented by a tree (dendogram)

Cut the tree at some level \( \Rightarrow \) obtain a number of clusters

Agglomerative approach: regard each data object as an individual cluster and then keep merging closest clusters until all merge into a single cluster

Measures of proximity of two clusters:

1. Single link - the shortest distance from any member of one cluster to any member of the other cluster
2. Complete link - the greatest distance from any member of one cluster to any member of the other cluster
3. Average link - the average distance between cluster members

Denote: \( D = \{d_{ij}\} \) - proximity matrix, \( L(k) \) - the \( k^{th} \) clustering level, \( (m) \) - the cluster formed in the \( m^{th} \) step, \( d[(r),(s)] \) - proximity between clusters \( (r) \) and \( (s) \)