Gene expression & Clustering
Determining gene function

• Sequence comparison tells us if a gene is similar to another gene, e.g., in a new species
  – Dynamic programming
  – Approximate pattern matching
• Genes with similar sequence likely to have similar function
• Doesn’t always work.
  – “Homologous” genes may not be similar enough at the sequence level, to be detected this way
• New method to determine gene function: directly measure gene activity (DNA arrays)
DNA Arrays--Technical Foundations

- An array works by exploiting the ability of a given mRNA molecule to hybridize to the DNA template.

- Using an array containing many DNA samples in an experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array.

- With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

May, 11, 2004

http://www.ncbi.nih.gov/About/primer/microarrays.html
An experiment on a microarray

In this schematic:

**GREEN** represents **Control DNA**

**RED** represents **Sample DNA**

**YELLOW** represents **a combination of Control and Sample DNA**

**BLACK** represents areas where **neither the Control nor Sample DNA**

Each color in an array represents either healthy (control) or diseased (sample) tissue. The location and intensity of a color tell us whether the gene is present in the control and/or sample DNA.
DNA Microarray

Millions of DNA strands build up on each location.

Tagged probes become hybridized to the DNA chip’s microarray.

May, 11, 2004

http://www.affymetrix.com/corporate/media/image_library/image_library_1.affx
Gene expression

• Microarray gives us an \( n \times m \) expression matrix \( I \)
  – Each of \( n \) rows corresponds to a gene
  – Each of \( m \) columns corresponds to a condition or time point
  – Each column comes from one microarray
• \( I(j,k) \) is the expression level of gene \( j \) in condition/experiment \( k \)
• If two genes (rows) have similar “expression profiles”, then
  – they may be related in function
  – they may be “co-regulated”
Clustering

• Find groups of genes that have similar expression profiles to one another
• Such groups may be functionally related, and/or co-regulated
• Compute pairwise distance metric $d(i,j)$ for every pair of genes $i$ and $j$
• This gives an $n \times n$ “distance matrix” $d$
Goal of clustering

• To group together genes into clusters such that
  – Genes within a cluster have highly similar expression profiles (small $d(i,j)$): “homogeneity”
  – Genes in different clusters have very different expression profiles (large $d(i,j)$): “separation”

• “Good” clustering is one that adheres to these goals

• A really “good” clustering is decided by biological interpretation of the clusters
Clustering of Microarray Data

(a) Intensity matrix, $I$

(b) Distance matrix, $d$

(c) Expression patterns as points in three-dimensional space.

Clusters
Clustering problems

- How to measure distance/similarity?
- How many clusters?
- Very large data sets: ~10,000 gene, ~100 conditions create computational difficulties
Hierarchical clustering

- One approach to clustering
- Does not explicitly partition genes into groups
- Organizes genes into a tree; genes are at the leaves of the tree
- Edges have lengths
- Total path length between two genes (leaves) correlates with the distance between the genes
Hierarchical Clustering
Hierarchical Clustering:
Hierarchical Clustering:
Hierarchical Clustering:
Hierarchical Clustering:

[Diagram showing hierarchical clustering with clusters labeled 1 to 10 and corresponding groupings.]
Hierarchical Clustering:
Hierarchical Clustering Algorithm

1. Hierarchical Clustering \((d, n)\)
2. Form \(n\) clusters each with one element
3. Construct a graph \(T\) by assigning one vertex to each cluster
4. while there is more than one cluster
5. Find the two closest clusters \(C_1\) and \(C_2\)
6. Merge \(C_1\) and \(C_2\) into new cluster \(C\) with \(|C_1| + |C_2|\) elements
7. Compute distance from \(C\) to all other clusters
8. Add a new vertex \(C\) to \(T\) and connect to vertices \(C_1\) and \(C_2\)
9. Remove rows and columns of \(d\) corresponding to \(C_1\) and \(C_2\)
10. Add a row and column to \(d\) corresponding to the new cluster \(C\)
11. return \(T\)

Different ways to define distances between clusters may lead to different clusterings
Hierarchical Clustering: Recomputing Distances

- \( d_{\text{min}}(C, C^*) = \min d(x,y) \)
  for all elements \( x \) in \( C \) and \( y \) in \( C^* \)

- Distance between two clusters is the **smallest**
  distance between any pair of their elements

- \( d_{\text{avg}}(C, C^*) = \frac{1}{|C^*||C|} \sum d(x,y) \)
  for all elements \( x \) in \( C \) and \( y \) in \( C^* \)

- Distance between two clusters is the **average**
  distance between all pairs of their elements
K-means clustering

- Another popular solution to the clustering problem
- Guess a number k, which is the number of clusters that will be reported
- Finds explicit clusters (unlike hierarchical clustering)
Objective function

- Let $V = (v_1, v_2, \ldots, v_n)$ be the $n$ data points
  - Each data point for is a vector in an $m$-dimensional space
- Let $X$ be a set of $k$ points in the same vector space
- For each $v_i$, find the $x \in X$ that is closest to it, i.e., the Euclidian distance $d(v_i, x)$ is least
- Sum the square of this Euclidian distance, over all $v_i$
- Formally,
  $$d(v_i, X) = \min_{x \in X} d(v_i, x)$$
  $$d(V, X) = \frac{\sum_{i} d(v_i, X)^2}{n}$$
Objective function

- Find $X$ such that $d(V,X)$ is minimized
- **Input**: A set, $V$, consisting of $n$ points and a parameter $k$
- **Output**: A set $X$ consisting of $k$ points (cluster centers) that minimizes the $d(V,X)$ over all possible choices of $X$
- An NP-complete problem for $k > 1$
  - Easy for $k=1$!
K-Means Clustering: Lloyd Algorithm

1. **Lloyd Algorithm**
2. Arbitrarily assign the \( k \) cluster centers
3. **while** the cluster centers keep changing
4. Assign each data point to the cluster \( C_i \) corresponding to the closest cluster representative (center) \( (1 \leq i \leq k) \)
5. After the assignment of all data points, compute new cluster representatives according to the center of gravity of each cluster, that is, the new cluster representative is
   \[ \frac{\sum v}{|C|} \text{ for all } v \text{ in } C \] for every cluster \( C \)

*This may lead to merely a locally optimal clustering.*
Conservative K-Means Algorithm

- Lloyd algorithm is fast but in each iteration it moves many data points, not necessarily causing better convergence.

- A more conservative method would be to move one point at a time only if it improves the overall clustering cost

  - The smaller the clustering cost of a partition of data points is, the better that clustering is

  - Different methods (e.g., \( d(V,X) \) we saw earlier) can be used to measure this clustering cost
K-Means “Greedy” Algorithm

1. \textbf{ProgressiveGreedyK-Means}(k)
2. Select an arbitrary partition \( P \) into \( k \) clusters
3. while forever
4. \hspace{1em} \textit{bestChange} \leftarrow 0
5. \hspace{1em} for every cluster \( C \)
6. \hspace{2em} for every element \( i \) not in \( C \)
7. \hspace{3em} if moving \( i \) to cluster \( C \) reduces its clustering cost
8. \hspace{1em} if (\text{cost}(P) – \text{cost}(P_i \rightarrow C) > \text{bestChange})
9. \hspace{2em} \textit{bestChange} \leftarrow \text{cost}(P) – \text{cost}(P_i \rightarrow C)
10. \hspace{3em} \textit{i}^* \leftarrow i
11. \hspace{3em} \textit{C}^* \leftarrow C
12. if \( \text{bestChange} > 0 \)
13. \hspace{1em} Change partition \( P \) by moving \( \textit{i}^* \) to \( \textit{C}^* \)
14. else
15. \hspace{1em} return \( P \)