Clustering Gene Expression Data

Harry Zuzan

Room 217 Old Chemistry Building

684-4447

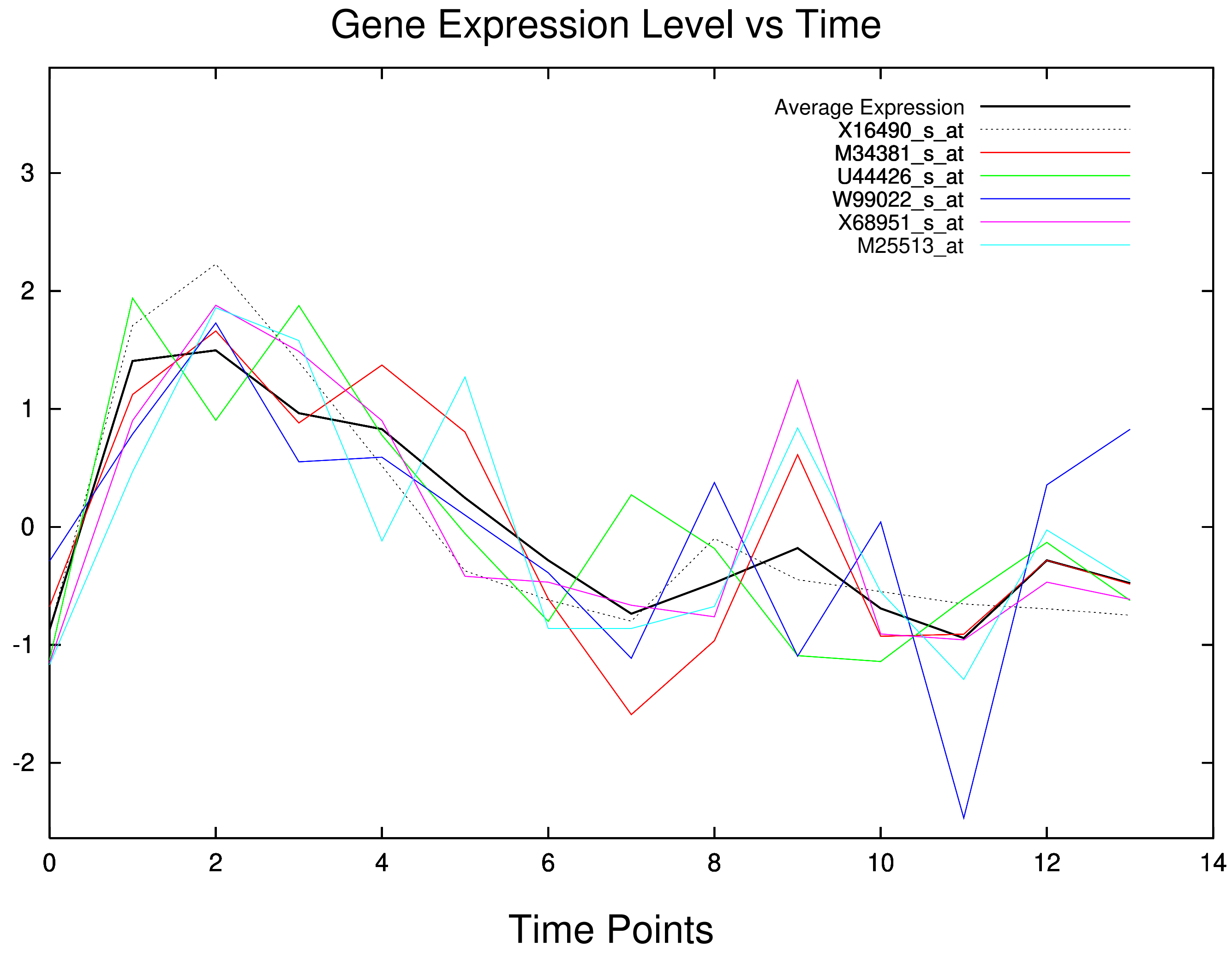
harry@stat.duke.edu

Cell Cycle Experiment

Identification of transcription factors regulating DNA replication and cell cycle progression.

1. Synchronize cells in a tissue culture.
2. Stimulate cell growth and take tissue samples over 14 time points spanning one cell cycle.
3. Extract RNA from the 14 samples and obtain measured expression levels for a large number of genes.
4. Look for similar expression expression patterns using clustering methods.

Gene Expression Patterns



Organizing Multivariate Data Into Matrices

Suppose p genes are measured once on each of n microarrays. Many programs which analyse these data wish to see them organized in a matrix in one of two ways.

1. Let each row correspond to a microarray and each column correspond to a gene.

2. Let each row correspond to a gene and each column correspond to a microarray.

Today we choose to use arrangement 2.

The Matrix of Observations

Let $x_{i,j}$ be the measurement of gene i from microarray j, where $i \in \{1, \cdots, p\}$ and $j \in \{1, \cdots, n\}$.

$$\mathbf{X} = [x_{i,j}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \ldots & x_{p,n} \end{bmatrix}$$

Interpreting Rows and Columns

The columns of $\mathbf{X}$ are "snapshots" of gene expression on each micorarray. Since the data for each column are from the same image, standardizing rows can be viewed as **image analysis**.

The rows of $\mathbf{X}$ can be used to reveal changes in expression levels of specific genes. Comparing rows to one another can be viewed as **pattern recognition**.

Quantifying Similarities and Dissimilarities

In order to perform a preliminary analysis, such as clustering, on the data contained in $\mathbf{X}$, the idea of similarity or dissimilarity must be quantified.

There are two sets of distances to be concerned with in $\mathbf{X}$:

1. The set of pairwise distances between columns can be used to identify similar gene expression "snapshots".

2. The set of pairwise distances between rows can identify genes with similar behaviour across differing experimental conditions.

Distance Measures

Let $\mathbf{x}_q$, $\mathbf{x}_r$ and $\mathbf{x}_s$ be any three rows of $\mathbf{X}$.

Let $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_r)$ be a function which generates a distance between any two rows $\mathbf{x}_r$ and $\mathbf{x}_q$ of $\mathbf{X}$.

Common choice of d is the Euclidean distance.

$$\mathrm{d}(\mathbf{x}_q, \mathbf{x}_r) = \sqrt{\sum_{\mathrm{j}=1}^{n} (x_{q,j} - x_{r,j})^2}$$

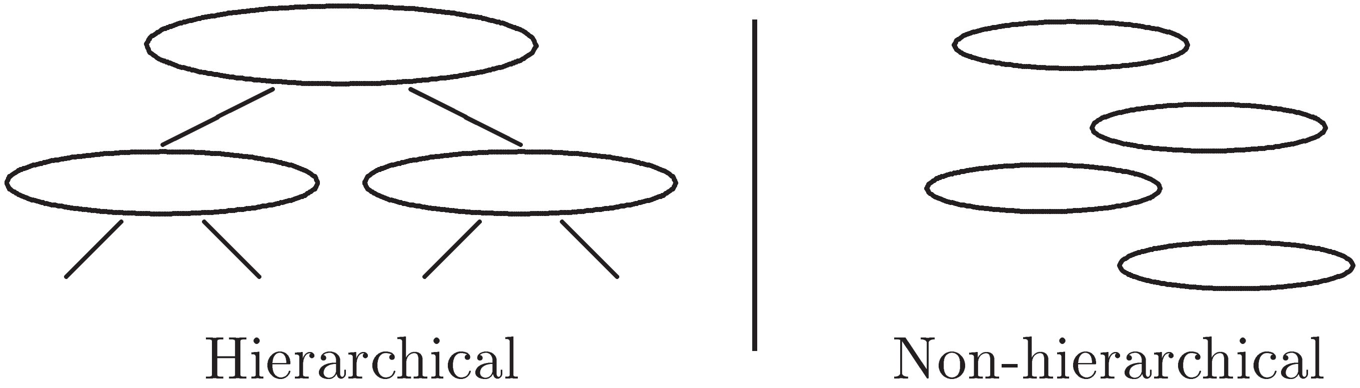
Distance Matrix

Writing $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_r)$ as $\mathrm{d}_{q,r}$,

$$\mathbf{D} = [\mathrm{d}(\mathbf{x}_q, \mathbf{x}_r)] = \begin{bmatrix} 0 & \mathrm{d}_{1,2} & \mathrm{d}_{1,3} & \ldots & \mathrm{d}_{1,p} \\ \mathrm{d}_{2,1} & 0 & \mathrm{d}_{2,3} & \ldots & \mathrm{d}_{2,p} \\ \mathrm{d}_{3,1} & \mathrm{d}_{3,2} & 0 & \ldots & \mathrm{d}_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{d}_{p,1} & \mathrm{d}_{p,2} & \mathrm{d}_{p,3} & \ldots & 0 \end{bmatrix}$$

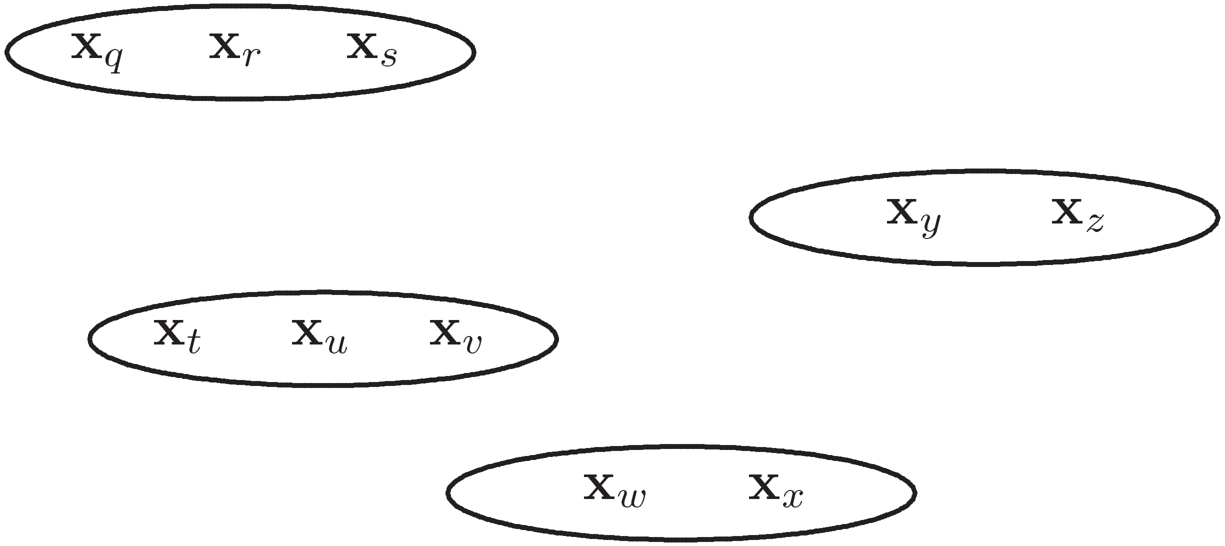
$\mathbf{D}$ is a symmetric matrix.

Clustering Methods



K-means Clustering (1)

K-means clustering partitions the data without imposing a hierarchical structure on the clusters.



K-means Clustering (2)

Clustering the p rows of $\mathbf{X}$ into K clusters.

1. Randomly assign each of $\mathbf{x}_1, \ldots, \mathbf{x}_p$, to the K clusters.
2. Compute the centroids $\mathbf{y}_1, \ldots \mathbf{y}_{\mathrm{K}}$, of the K clusters.
3. In the order $i = 1, \ldots, p$, compute the distance between $\mathbf{x}_i$ and each of the centroids $\mathbf{y}_1, \ldots \mathbf{y}_{\mathrm{K}}$. Assign $\mathbf{x}_i$ to the cluster with the closest centroid.
4. If $\mathbf{x}_i$ was moved to a new cluster update the centroids of both clusters affected.
5. Return to step 3 until no more reassignments will take place.

K-means Clustering (3)

For $k = 1, \ldots, \mathrm{K}$, let C_k be the set of all j such that column $\mathbf{x}_j$ is in cluster k.

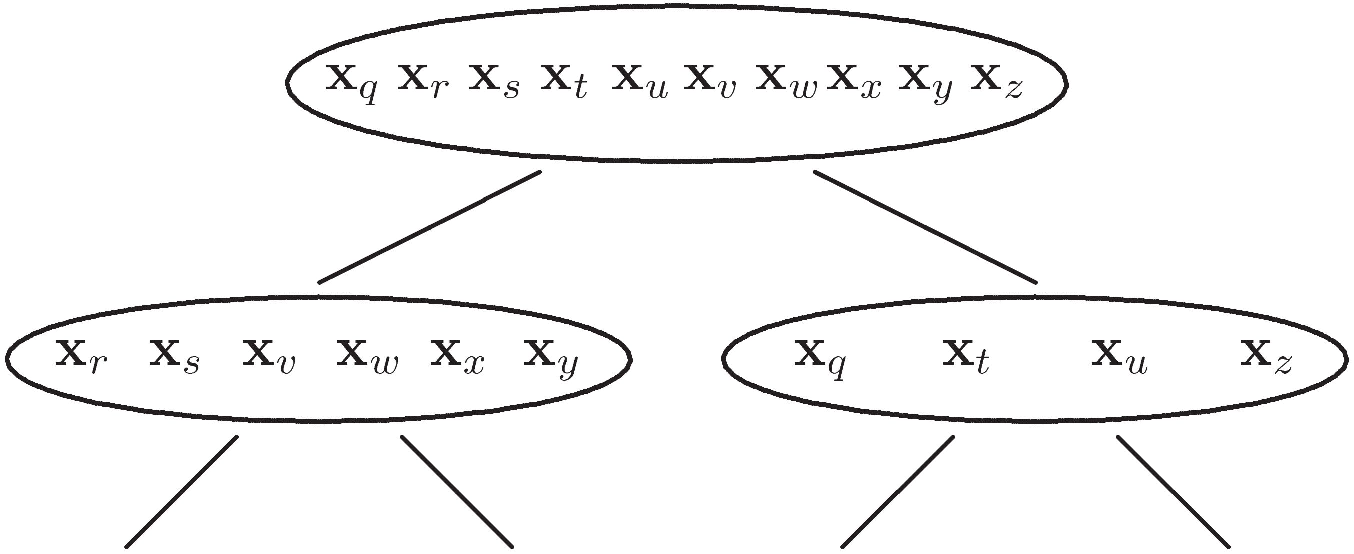
K-means clustering tries to find the sets $\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{K}$ which minimize the quantity

$$\sum_{k=1}^{\mathrm{K}} \sum_{j \in \mathrm{C}_k} ||\mathbf{x}_j - \mathbf{y}_k||^2$$

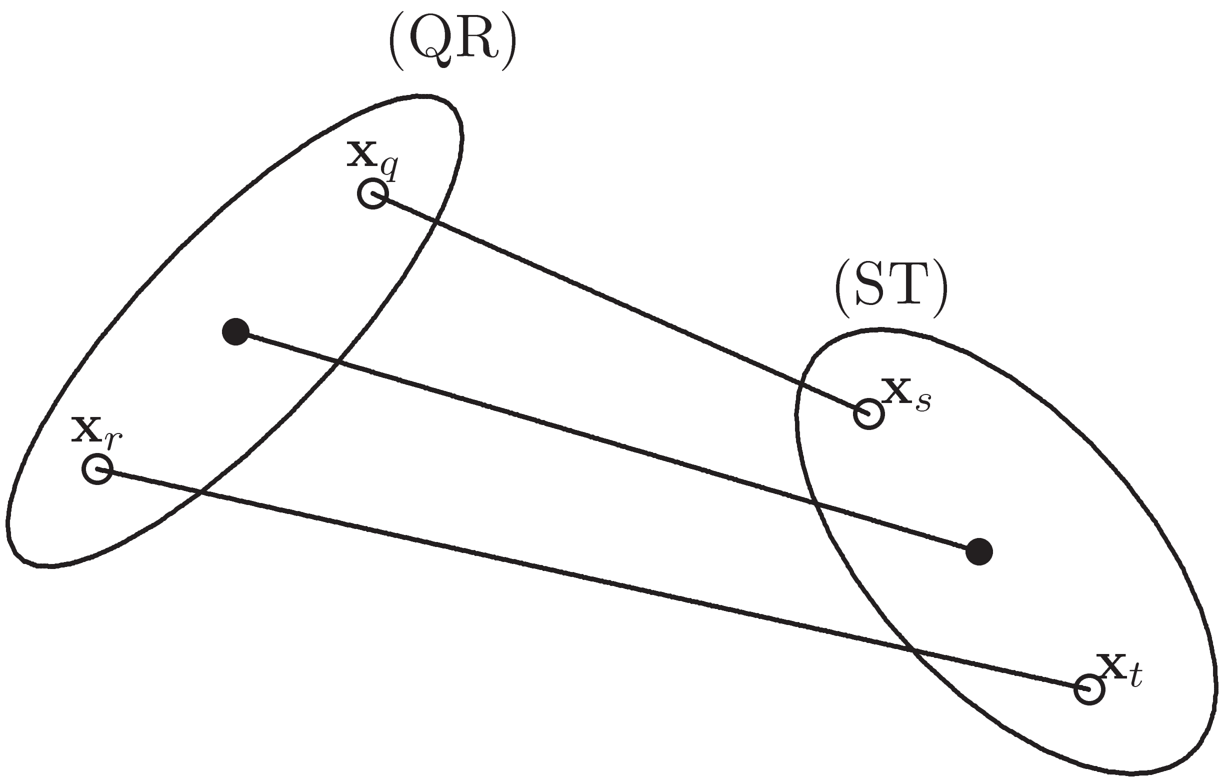
which is the sum of squares within clusters. An exhaustive search usually is not possible, even for small data sets.

Hierarchical Clustering

In hierarchical clustering, the data is divided by a binary tree.



Cluster Aggregration



Agglomerative Hierarchical Clustering (1)

Suppose the p rows of $\mathbf{X}$, $[\mathbf{x}_1, \ldots, \mathbf{x}_p]$, are to be clustered.

1. Start with p clusters, each containing one row of $\mathbf{X}$. And compute the $p \times p$ symmetric matrix of distances between the clusters.

2. Search the distance matrix for the **smallest distance between clusters**. Suppose the smallest distance is between $\mathbf{x}_q$ and $\mathbf{x}_r$. Label these clusters (Q) and (R).

Agglomerative Hierarchical Clustering (2)

3. Merge clusters (Q) and (R). Label the newly formed cluster (QR). Update the entries in the distance matrix by:
   (a) Deleting the rows and columns corresponding to clusters (Q) and (R).
   (b) Adding in a row and column giving the distances between cluster (QR) and the remaining clusters.
4. Repeat steps 2 and 3 a total of $p-1$ times. All of the rows of $\mathbf{X}$ will then be found in a single cluster after $p-1$ iterations.

Agglomerative Hierarchical Clustering (3)

- The meaning of **smallest distance between clusters** is determined by the type of clustering algorithm, such as: Single linkage, complete linkage or average linkage.

Single Linkage Hierarchical Clustering

The distance between two clusters is the distance between **nearest neighbours**.

Suppose cluster (QR) contains $\mathbf{x}_q$ and $\mathbf{x}_r$, and cluster (ST) contains $\mathbf{x}_s$ and $\mathbf{x}_t$. The distance between (QR) and (ST) will be the **minimum** of $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_s)$, $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_t)$, $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_s)$, and $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_t)$.

Single linkage is useful when clusters are well defined but oddly shaped. The single linkage algorithm is also invariant to monotonic transformations of distance d.

Complete Linkage Hierarchical Clustering

The distance between two clusters is the largest distance between any pair of data not in the same cluster.

Suppose cluster (QR) containing $\mathbf{x}_q$ and $\mathbf{x}_r$, and cluster (ST) containing $\mathbf{x}_s$ and $\mathbf{x}_t$. The distance between (QR) and (ST) is the **maximum** of $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_s)$, $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_t)$, $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_s)$, and $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_t)$.

Complete linkage ensures that all data within any cluster are within some maximum distance.

Average Linkage Hierarchical Clustering

The distance between two clusters is the average distance between all pairs of data in opposing clusters.

Suppose cluster (QR) contains $\mathbf{x}_q$ and $\mathbf{x}_r$, and cluster (ST) contains $\mathbf{x}_s$ and $\mathbf{x}_t$. The distance between (QR) and (ST) is the **average** of $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_s)$, $\mathrm{d}(\mathbf{x}_q, \mathbf{x}_t)$, $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_s)$, and $\mathrm{d}(\mathbf{x}_r, \mathbf{x}_t)$.

Hierarchical / Non-hierarchical

Hierarchical Clusters

- The same result will be got each time with the same data.
- Can be difficult to decide on a representative member for each cluster. Especially with single linkage.

K-means Clustering

- The final result depends on the initial partition of the data.
- The centroid of each cluster is a natural representation, or summary of its membership.