Clustering methods

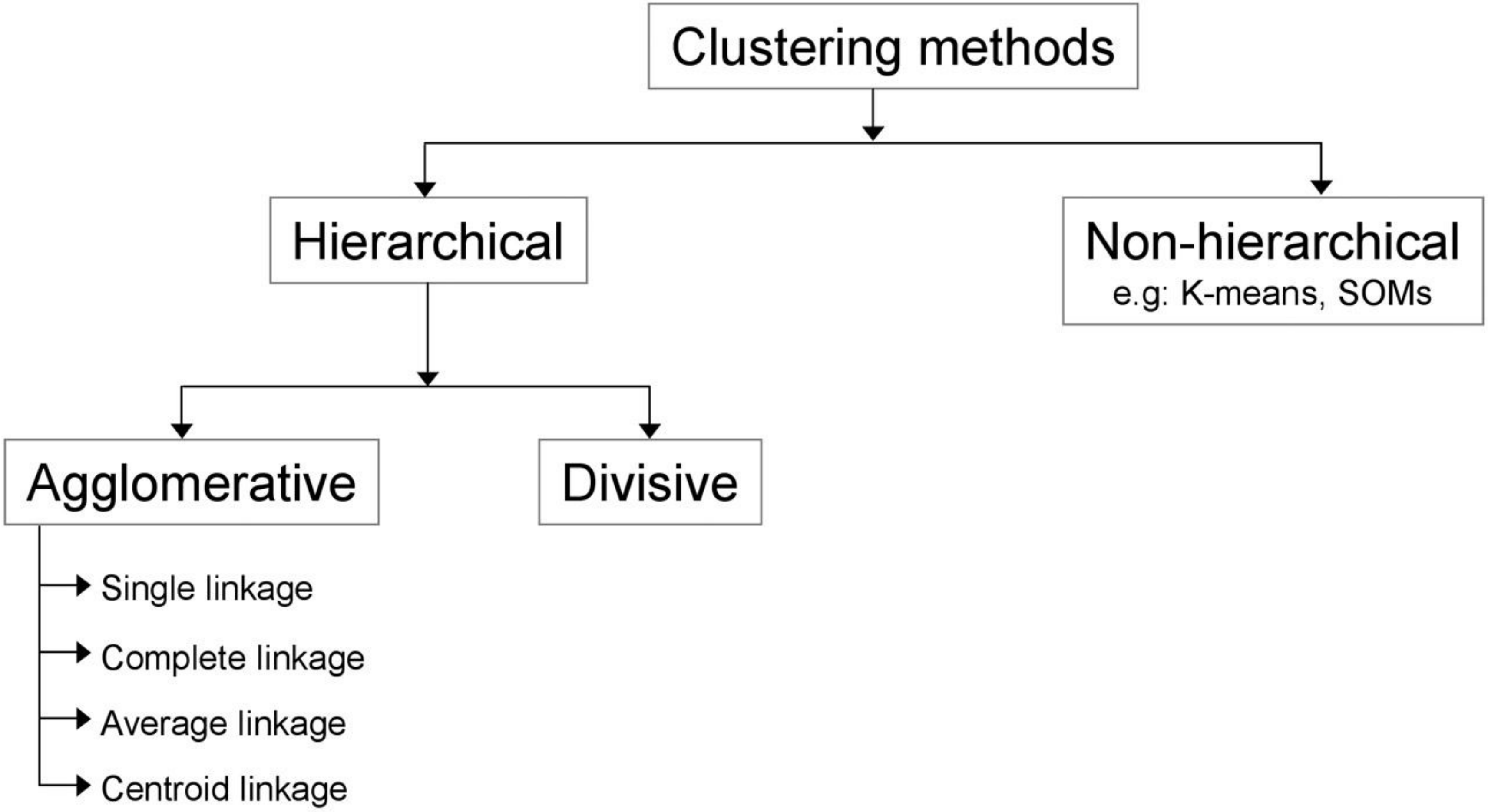
MBI401-High throughput Data Generation & Analysis

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

- One of the goals of microarray data analysis is to cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples.
- Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group

- Clustering methods can be hierarchical (grouping objects into clusters and specifying relationships among objects in a cluster, resembling a phylogenetic tree) or
- non-hierarchical (grouping into clusters without specifying relationships between objects in a cluster) as schematically represented in Figure 5. Remember, an object may refer to a gene or a sample, and a cluster refers to a set of objects that behave in a similar manner

Figure 5. An overview of the different clustering methods.



Clustering methods

Hierarchical

Non-hierarchical
e.g: K-means, SOMs

Agglomerative

Divisive

Single linkage

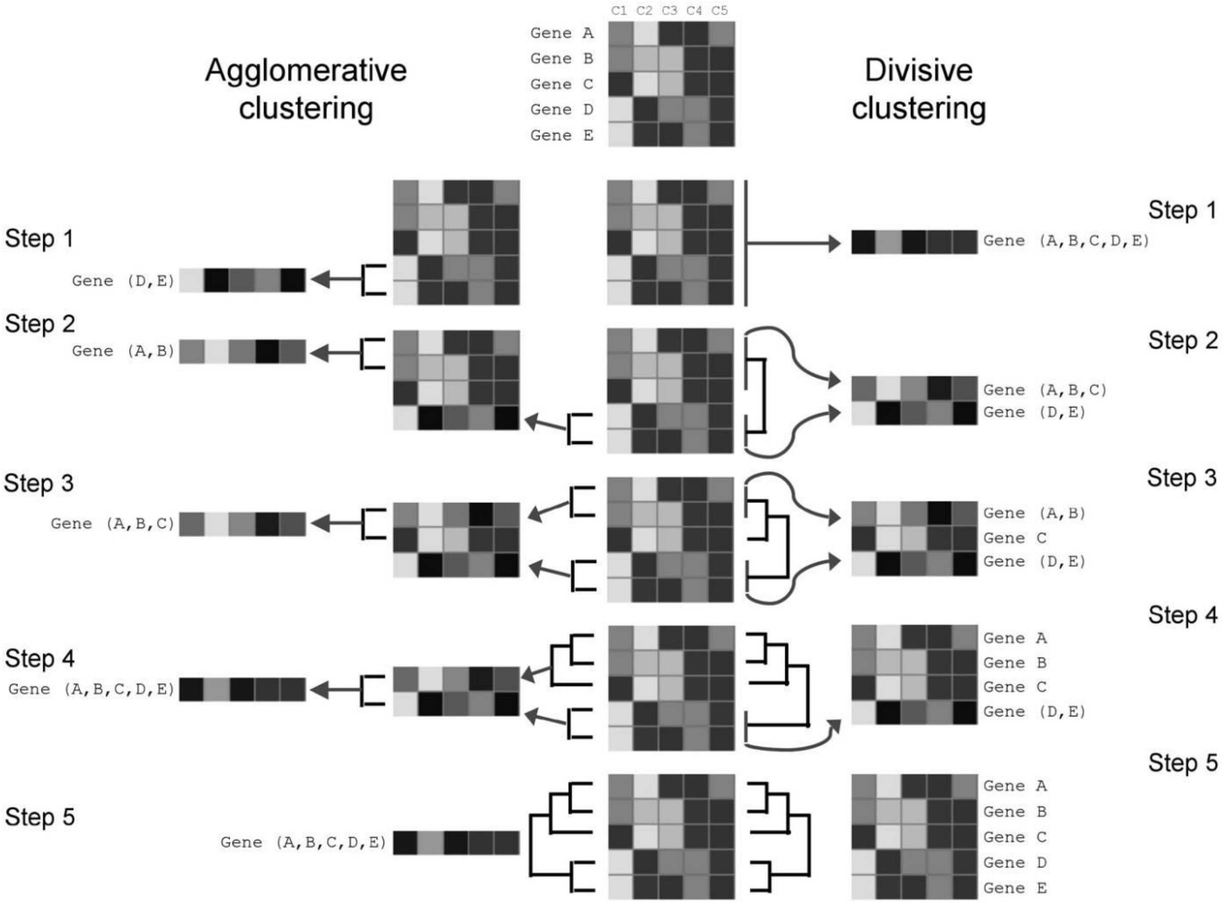
Complete linkage

Average linkage

Centroid linkage

- ***Hierarchical clustering***
- Hierarchical clustering may be agglomerative (starting with the assumption that each
- object is a cluster and grouping similar objects into bigger clusters) or divisive (starting
- from grouping all objects into one cluster and subsequently breaking the big cluster into
- smaller clusters with similar properties).

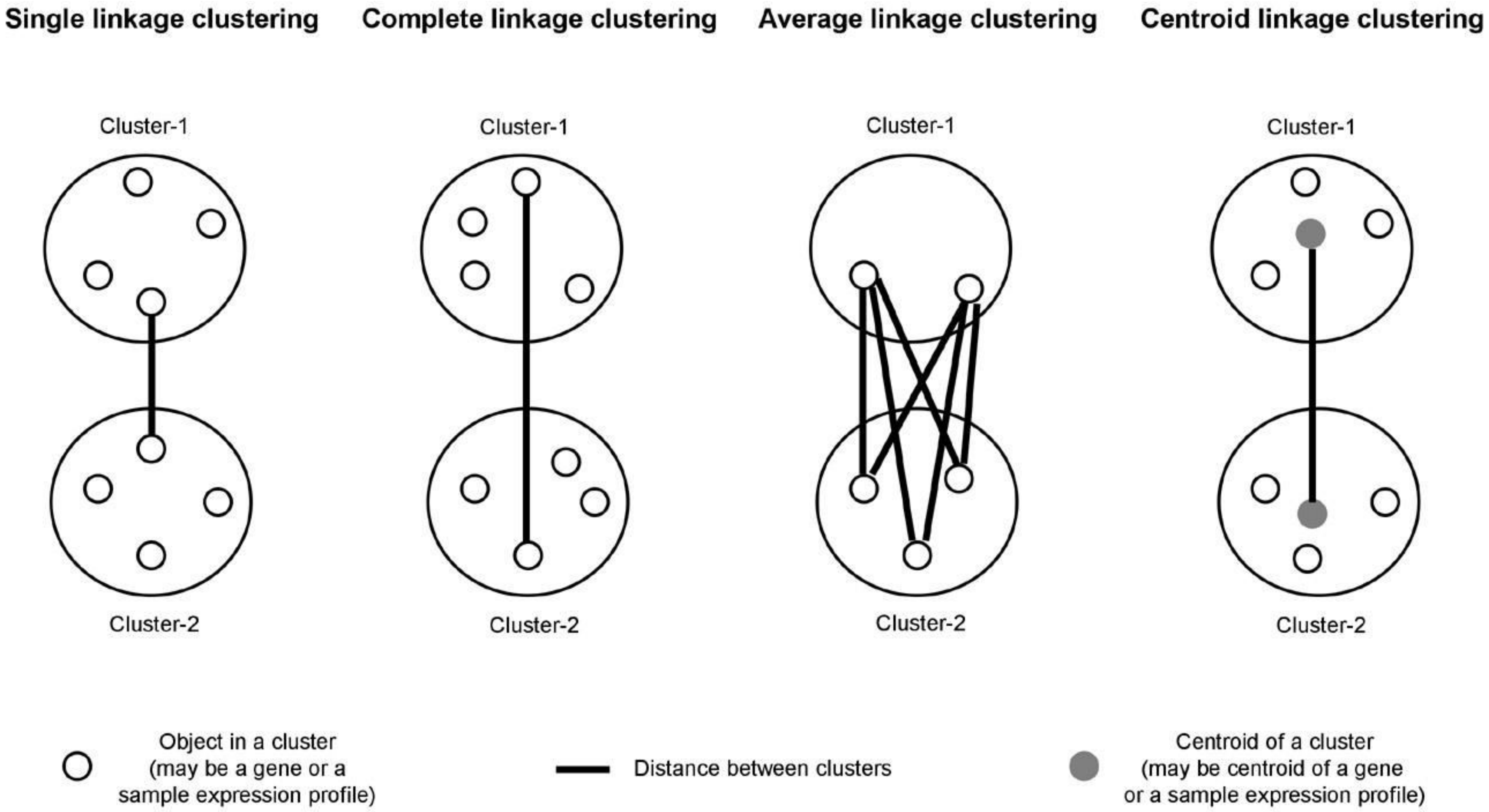
Figure 6. Schematic diagram showing the principle behind agglomerative and divisive clustering. The colour code represents the log2 (expression ratio), where red represents up-regulation, green represents down-regulation, and black represents no change in expression



Hierarchical clustering: agglomerative

- In the case of a hierarchical agglomerative clustering, the objects are successively fused
- until all the objects are included. For a hierarchical agglomerative clustering procedure,
- each object is considered as a cluster. The first step is the calculation of pairwise distance
- measures for the objects to be clustered. Based on the pairwise distances between them,
- objects that are similar to each other are grouped into clusters. After this is done, pairwise
- distances between the clusters are re-calculated, and clusters that are similar are grouped
- together in an iterative manner until all the objects are included into a single cluster

Figure 7. Different algorithms to find distance between two clusters.


Single linkage clustering
Complete linkage clustering
Average linkage clustering
Centroid linkage clustering
Cluster-1
Cluster-2
Cluster-1
Cluster-2
Cluster-1
Cluster-2
Cluster-1
Cluster-2
Object in a cluster (may be a gene or a sample expression profile)
Distance between clusters
Centroid of a cluster (may be centroid of a gene or a sample expression profile)

Single linkage clustering (Minimum distance)

- In single linkage clustering, distance between two clusters is calculated as the minimum
- distance between all possible pairs of objects, one from each cluster. This method has
- an advantage that it is insensitive to outliers. This method is also known as the nearest
- neighbour linkage.

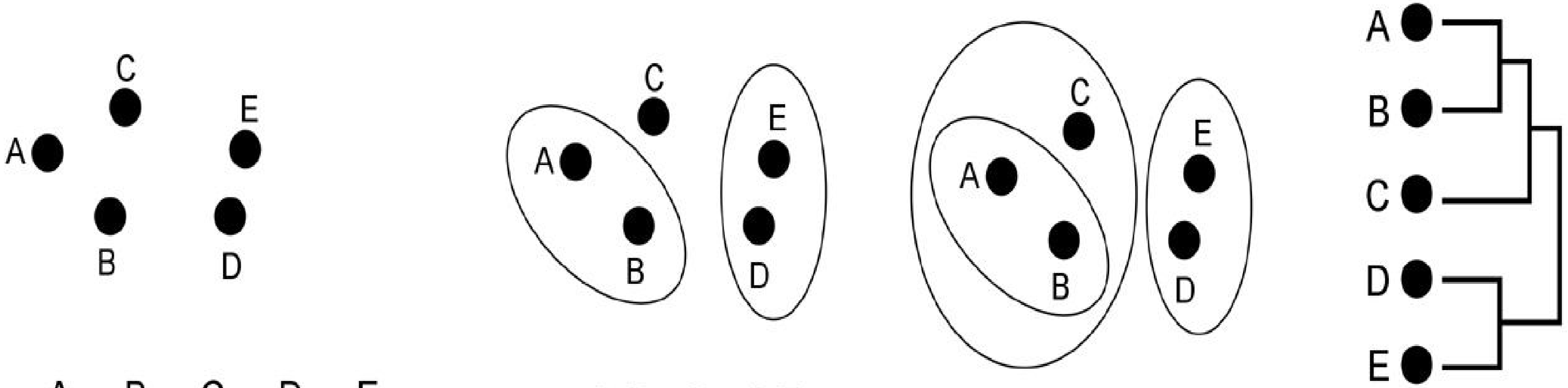
Complete linkage clustering (Maximum distance)

- In complete linkage clustering, distance between two clusters is calculated as the maximum.
- distance between all possible pairs of objects, one from each cluster. The disadvantage
- of this method is that it is sensitive to outliers. This method is also known as the farthest
- neighbour linkage.

Average linkage clustering

- In average linkage clustering, distance between two clusters is calculated as the average of
- distances between all possible pairs of objects in the two clusters.

Figure 8. An example of a hierarchical clustering using single linkage algorithm. Consider five genes and the distances between them as shown in the table. In the first step, genes that are close to each other are grouped together and the distances are re-calculated using the single linkage algorithm. This procedure is repeated until all genes are grouped into one cluster. This information can be represented as a tree (shown to the right), where the distance from the branch point reflects the distance between genes or clusters. This image was adapted from Causton *et al. (2003).*



	A	B	C	D	E
A	0	1	2	3	4
B		0	3	4	5
C			0	4	5
D				0	1
E					0

	A,B	C	D,E
A,B	0	2	3
C		0	4
D,E			0

	A,B,C	D,E
A,B,C	0	3
D,E		0

Centroid linkage clustering

- In centroid linkage clustering, an average expression profi le (called a centroid) is calculated in two steps. First, the mean in each dimension of the expression profi les is calculated for all objects in a cluster. Then, distance between the clusters is measured as the distance
- between the average expression profi les of the two clusters.
- An example of the hierarchical agglomerative clustering using single linkage clustering
- is shown in Figure 8.

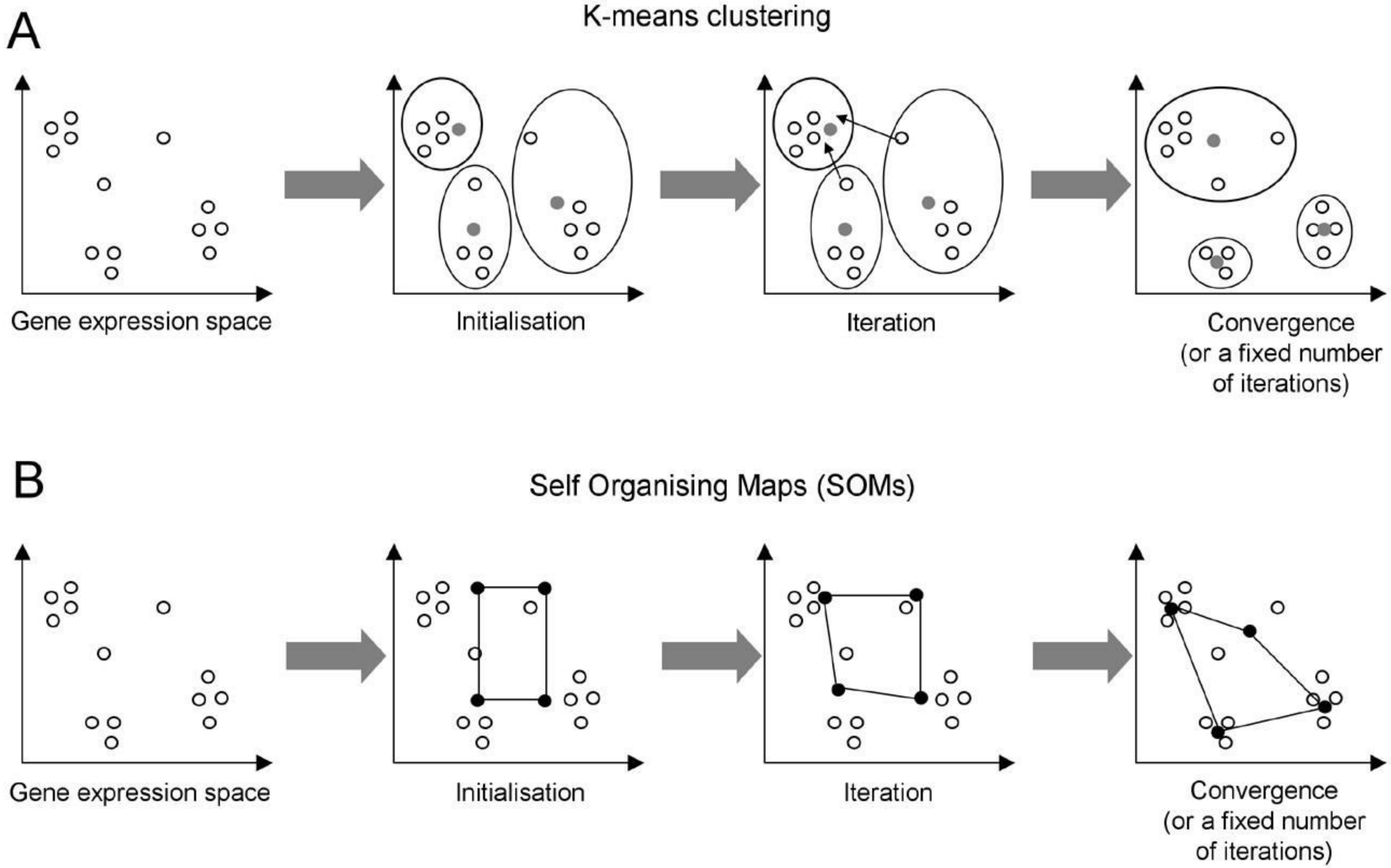
Hierarchical clustering: divisive

- Hierarchical divisive clustering is the opposite of the agglomerative method, where the entire set of objects is considered as a single cluster and is broken down into two or more clusters that have similar expression profiles.
- After this is done, each cluster is considered separately and the divisive process is repeated iteratively until all objects have been separated into single objects.
- The division of objects into clusters on each iterative step may be decided upon by principal component analysis which determines a vector that separates given objects.
- This method is less popular than agglomerative clustering, but has successfully been used in the analysis of gene expression data by Alon *et al. (1999).*

Non-hierarchical clustering

- One of the major criticisms of hierarchical clustering is that there is no compelling evidence that a hierarchical structure best suits grouping of the expression profiles.
- An alternative to this method is a non-hierarchical clustering, which requires predetermination of the number of clusters.
- Non-hierarchical clustering then groups existing objects into these predefined clusters rather than organizing them into a hierarchical structure.

Figure 9. A: The principle behind K-means clustering. Objects are grouped into a predefined number of clusters during the initialization step. Centroid for each cluster is calculated, and objects are re-grouped depending on how close they are to available centroids. This step is performed iteratively until convergence or is performed for a fixed number of iterations to get final clusters of objects. B: The principle behind SOMs. During the initialization step, a grid of nodes is projected onto the expression space and each gene is assigned its closest node. Following this step, one gene is chosen at random and the assigned node is 'moved' towards it. The other nodes are moved towards this gene depending on how close they are to the selected gene. This step is performed iteratively until convergence or is performed for a fixed number of iterations to get a final map of nodes.



Non-hierarchical clustering: K-means

- K-means is a popular non-hierarchical clustering method (Figure 9A). In K-means clustering, the first step is to arbitrarily group objects into a predetermined number of clusters.
- The number of clusters can be chosen randomly or estimated by first performing a hierarchical clustering of the data.
- Following this step, an average expression profi le (centroid) is calculated for each cluster, this is called initialization. Next, individual objects are reattributed from one cluster to the other depending on which centroid is closer to the gene (or sample).

- This procedure of calculating the centroid for each cluster and re-grouping objects closer to available centroids is performed in an iterative manner for a fixed number of times, or until convergence (state when composition of clusters remains unaltered by further iterations).
- Typically, the number of iterations required to obtain stable clusters ranges from 20,000 to 100,000. However, there is no guarantee that the clusters will converge.
- This method has an advantage that it is scalable for large datasets.

- ***Non-hierarchical clustering: Self Organizing Maps***
- Self Organizing Maps (SOMs) work in a manner similar to K-means clustering (Figure 9B).
- In K-means clustering, one chooses the number of clusters to fit the data, whereas with SOM the first step is to choose the number and orientation of the clusters with respect to each other.
- For example, a two-dimensional grid of ‘nodes’ (which may end up being clusters)

- could be the starting point. The grid is projected onto the expression space, and each object is assigned a node that is nearest to it – this is called initialization.
- In the next step, a random object is chosen and the node (called a reference vector) which is in the ‘neighbourhood’ of the object is moved closer to it. The other nodes are moved to a small extent depending on how close they are to the object chosen.
- In successive iterations, with randomly chosen objects, the positions of the nodes are refined and the ‘radius of neighbourhood’ becomes confined. In this way, the grid of nodes (initially a two-dimensional grid) is deformed to fit the data.

- **4.1 Predicting binding sites**The steps involved in such studies are the following:
- (1) sequencesFind a set of genes that have similar expression profiles. (2) Extract promoter
- of the co-expressed genes. (3) Identify statistically over-represented sequence patterns. (4)
- Assess quality of the discovered pattern using statistical significance criteria.

4.2 Predicting protein interactions and protein functions

- The steps involved in such studies are the following:
- (1) Identify co-expressed genes in the two studied organisms.
- (2) Identify conserved (orthologous) proteins.
- (3) Find instances where conserved (orthologous) proteins are co-expressed in both organisms.
- (4) Map information on protein interaction or metabolic pathway available for one organism to predict interacting proteins or function of the proteins in the other organism.

4.3 Predicting functionally conserved modules

- Genes that have similar expression profiles often have related functions. Instead of studying co-expressed pairs of genes, one can view sets of co-expressed genes that are known to interact as a functional module involved in a particular biological process.
- The steps involved in such studies are similar to those discussed in the previous section.
- Instead of two organisms, one has to consider three or more organisms, and should also address other issues related to identifying orthologous proteins.

4.4 ‘Reverse-engineering’ of gene regulatory networks

- Gene expression data can also be used to infer regulatory relationships. This approach is known as reverse engineering of regulatory networks.
- Research by Segal *et al. (2003) and* Gardner *et al. (2003) clearly highlights that we are now in a good position to use expression* data to make predictions about the transcriptional regulators for a given gene or sets of genes. Segal *et al. (2003) have developed a probabilistic model to identify modules of coregulated* genes, their transcriptional regulators and conditions that influence regulation. This new knowledge allowed them to generate further hypotheses, which are experimentally testable.
- Gardner *et al. (2003) described a method to infer regulatory relationships, called* NIR (Network Identification by multiple Regression), which uses non-linear differential equations to model regulatory networks. In this method, a model of connections between genes in a network is inferred from measurements of system dynamics (*i.e. response of* genes and proteins to perturbations).

- **5.1 Website references**
- Some websites that provide a reference to various aspects of microarrays are given below:
- Portals:
- http://ihome.cuhk.edu.hk/%7Eb400559/array.html
- *A comprehensive web portal on microarrays.*
- http://www.bioinformatics.vg/biolinks/bioinformatics/Microarrays.shtml
- *A web-portal on microarrays.*
- http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-genexp.html
- *A collection of gene expression and microarray links at the HGMP (Human Genome Mapping Project).*