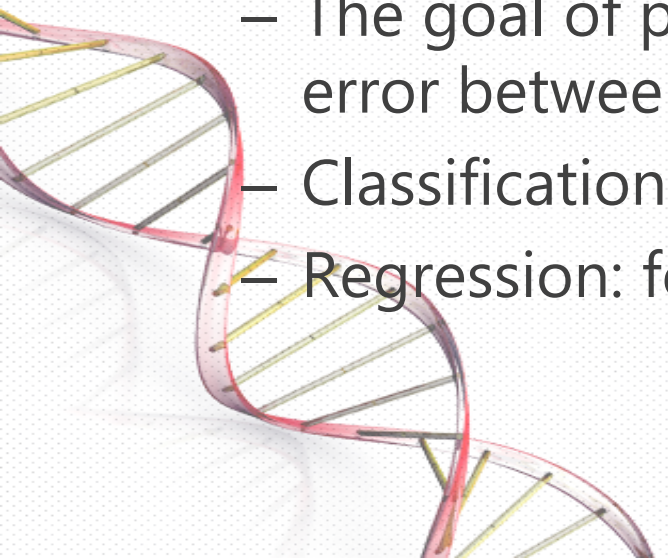
Clustering Algorithms

Hammad Naveed
hammad.naveed@nu.edu.pk

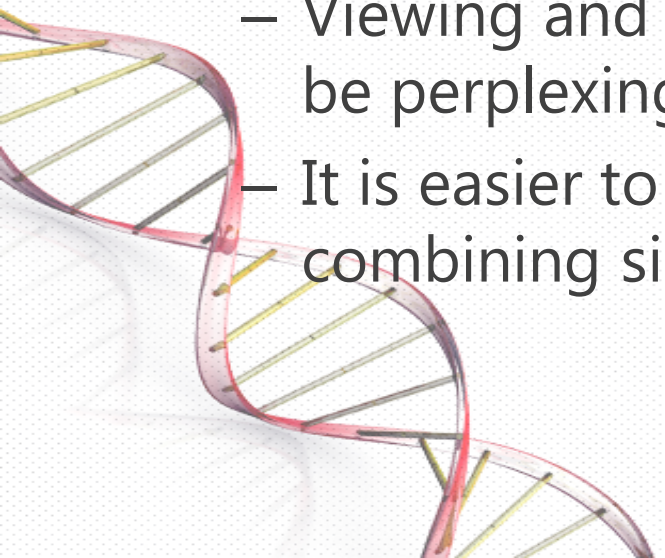
Data mining tasks

- Predictive vs descriptive tasks
  - Predictive tasks: predict the value of a particular attribute (the target variable) based on the values of other attributes
  - Descriptive tasks: derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data
- Predictive modeling
  - The goal of predictive modeling is to learn a model that minimizes the error between the predicted and true values of the target variable
  - Classification: for discrete target variables
  - Regression: for continuous target variables



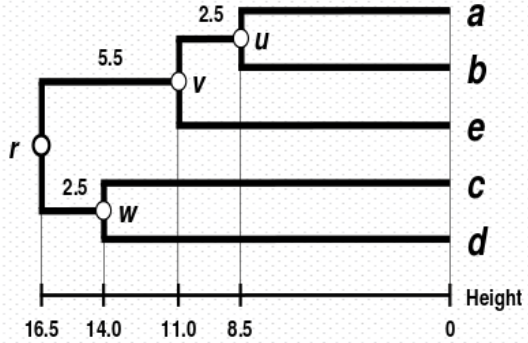
Data mining tasks

- Cluster analysis
  - The goal is to find groups of similar observations/objects
  - Many clustering algorithms have been developed
    - K-means
    - Hierarchical
- Applications of clustering in biology
  - Viewing and analyzing vast amounts of biological data as a whole set can be perplexing
  - It is easier to interpret the data if they are partitioned into clusters combining similar data points

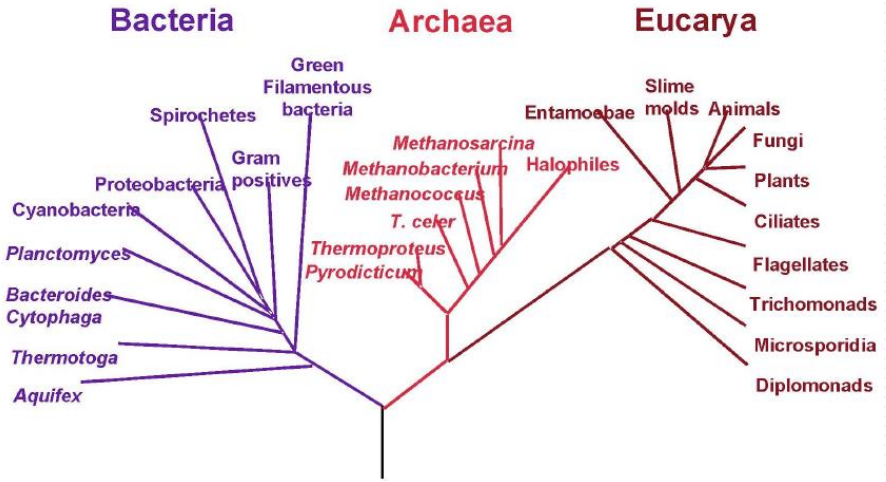


Clustering of genes/proteins

- **UPGMA** (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean) is applied to build guide tree for multiple sequence alignments

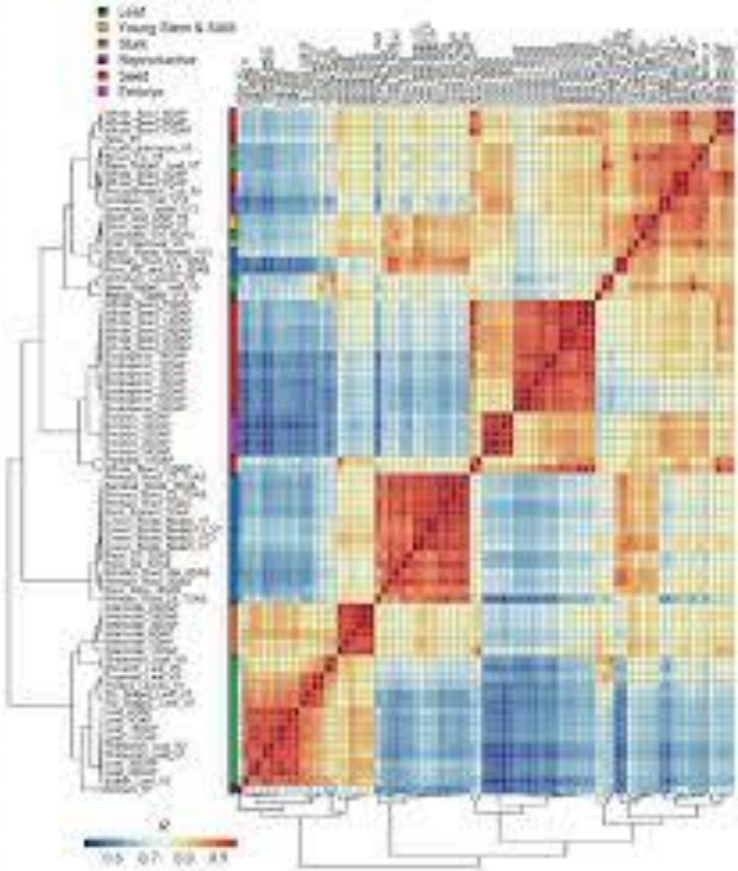


- Distance-based methods for phylogenetic reconstruction



Clustering of microarray data

- Plot each datum as a point in N-dimensional space
- Make a distance matrix for the distance between every two gene points in the N-dimensional space
- Genes with a small distance share the same expression characteristics and might be functionally related or similar.
- Clustering reveal groups of functionally related genes



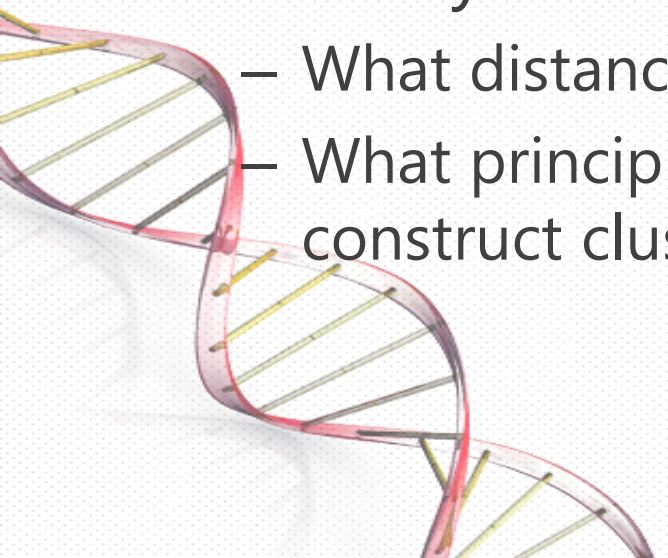
Clustering of microarray data

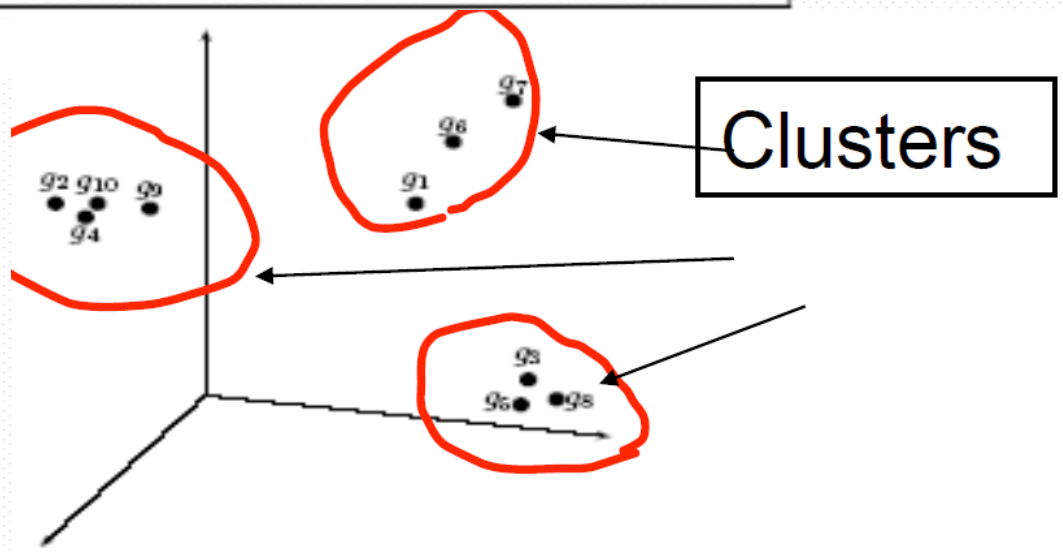
Time	1 hr	2 hr	3 hr
g_1	10.0	8.0	10.0
g_2	10.0	0.0	9.0
g_3	4.0	8.5	3.0
g_4	9.5	0.5	8.5
g_5	4.5	8.5	2.5
g_6	10.5	9.0	12.0
g_7	5.0	8.5	11.0
g_8	2.7	8.7	2.0
g_9	9.7	2.0	9.0
g_{10}	10.2	1.0	9.2

(a) Intensity matrix, I

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

- Two key factors:
  - What distance measure is used
  - What principle is used to construct clusters

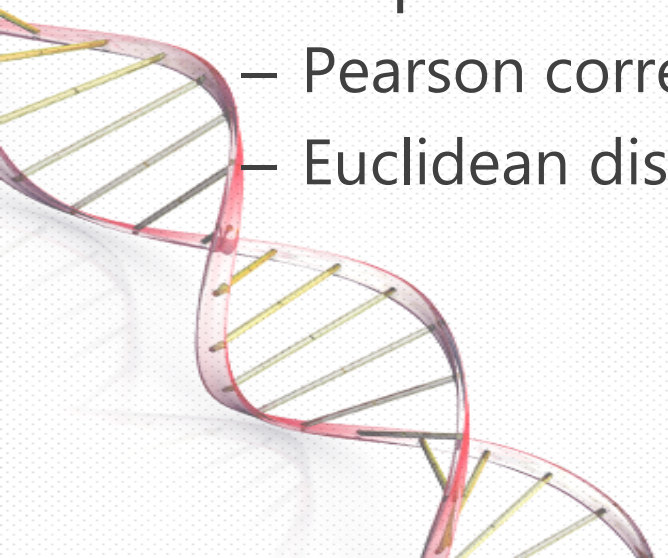




(c) Expression patterns as points in three-dimentsional space.

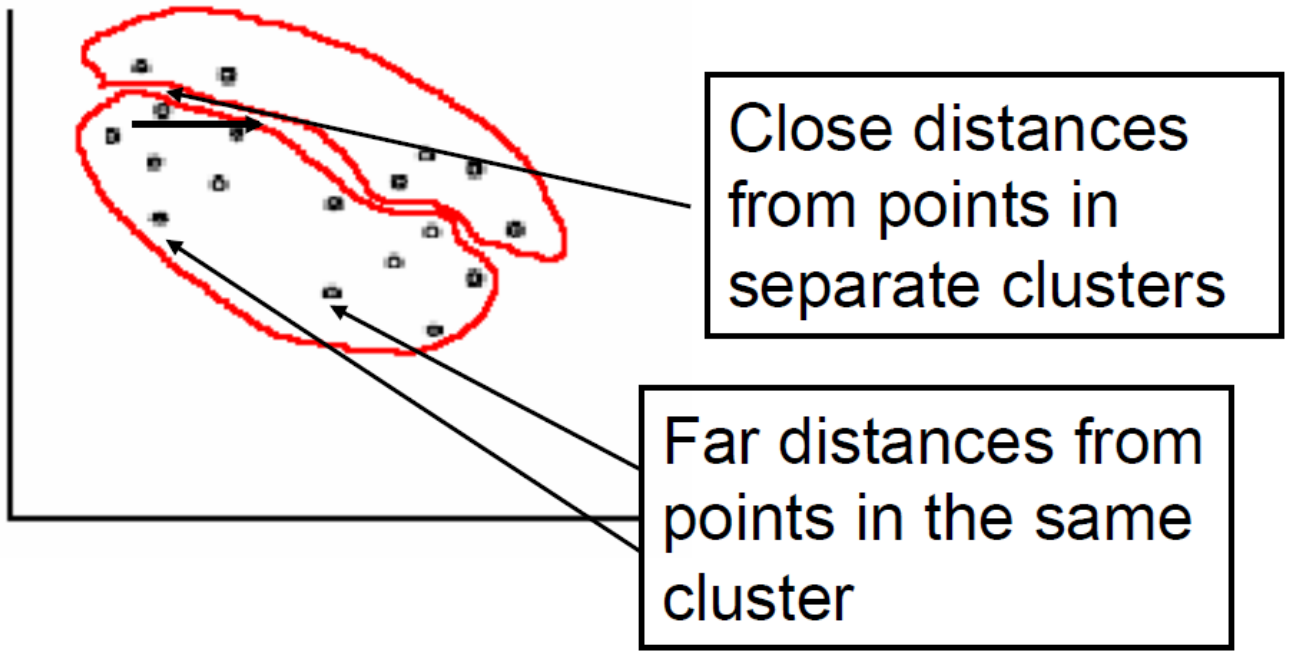
Measures of similarity and dissimilarity (distance)

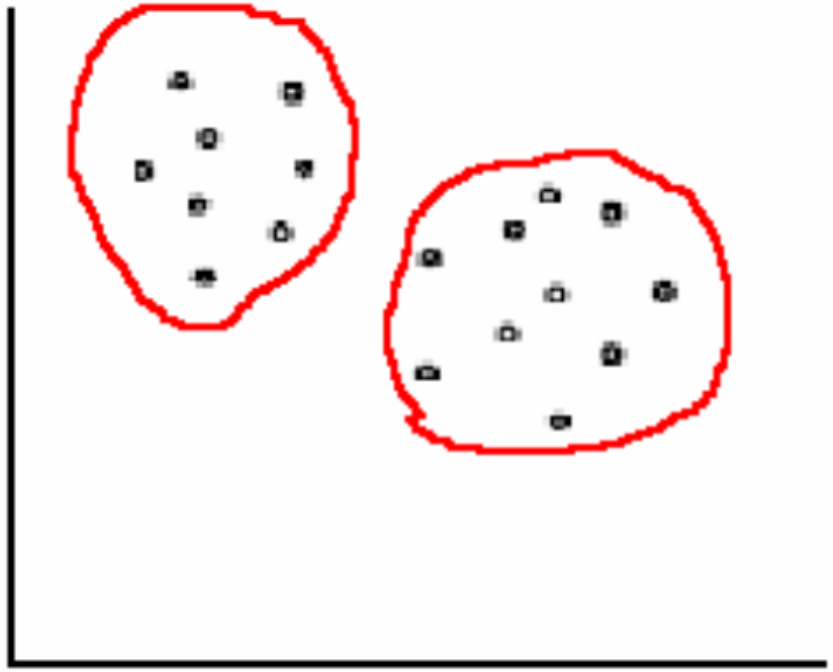
- There are many different ways of calculating similarity and distance
- Knowing your data is important
- When working on distance, pay attention to three properties: positivity, symmetry, and triangle inequality.
- Examples
  - Pearson correlation coefficient
  - Euclidean distance



Homogeneity and separation principles

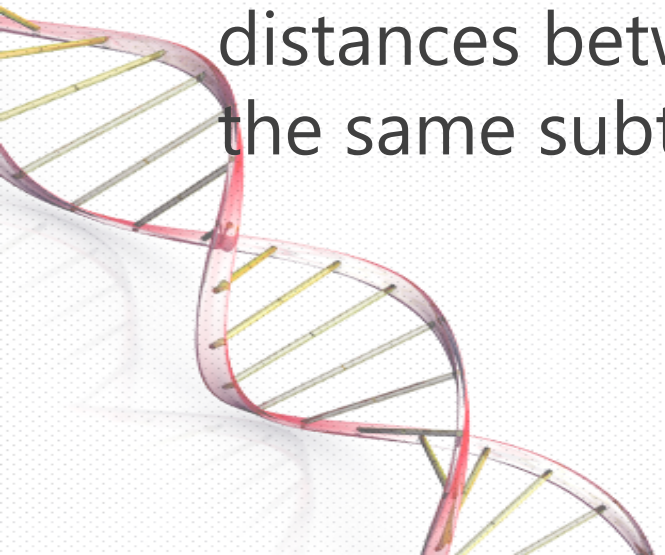
- **Homogeneity**: Elements within a cluster are close to each other
- **Separation**: Elements in different clusters are further apart from each other
- …clustering is not an easy task!



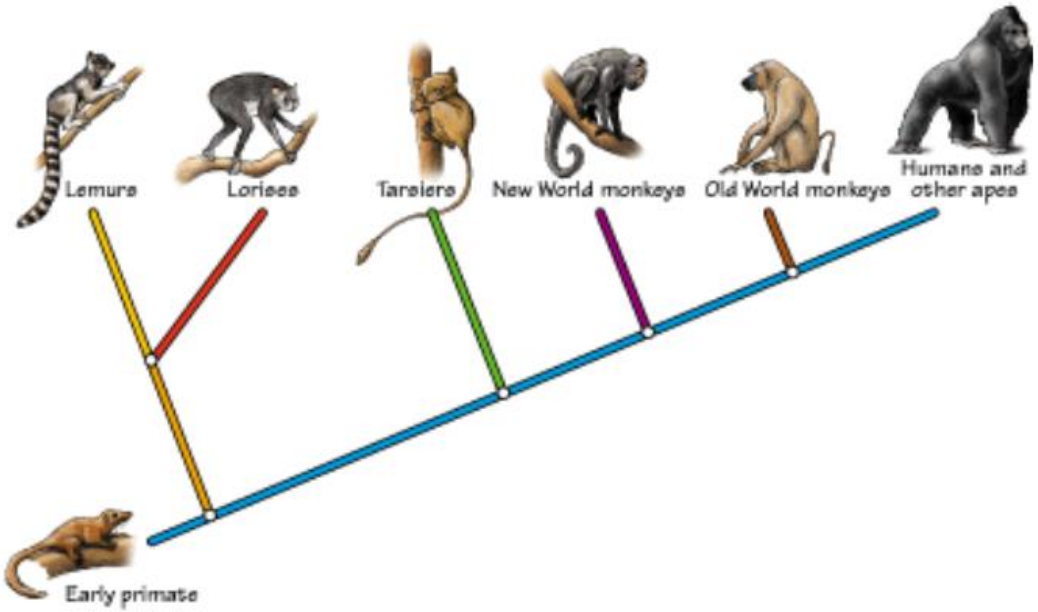


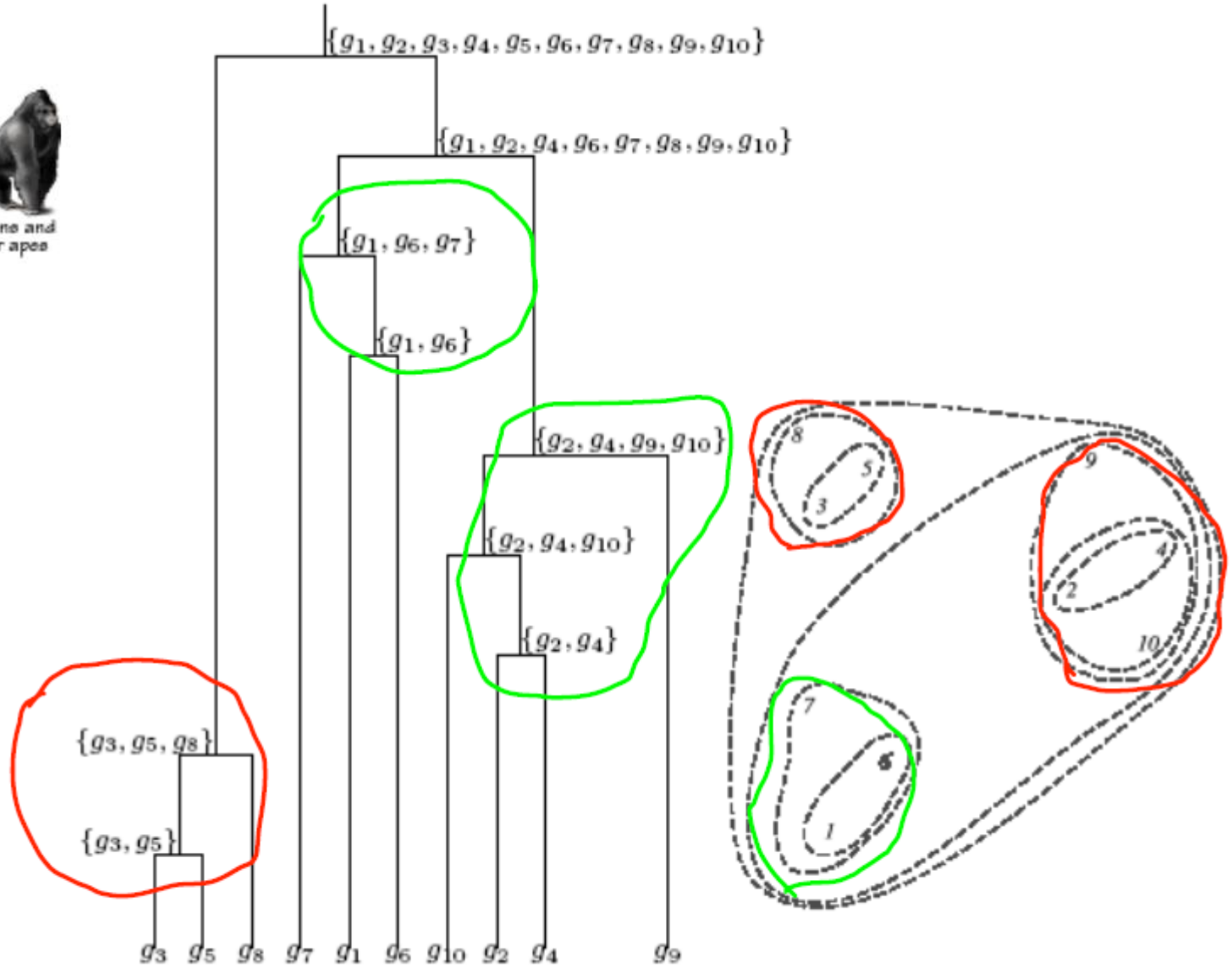
Clustering techniques

- **Agglomerative**: Start with every element in its own cluster, and iteratively join clusters together
- **Divisive**: Start with one cluster and iteratively divide it into smaller clusters
- **Hierarchical**: Organize elements into a tree, leaves represent genes and the length of the paths between leaves represents the distances between objects (genes, etc). Similar objects lie within the same subtrees



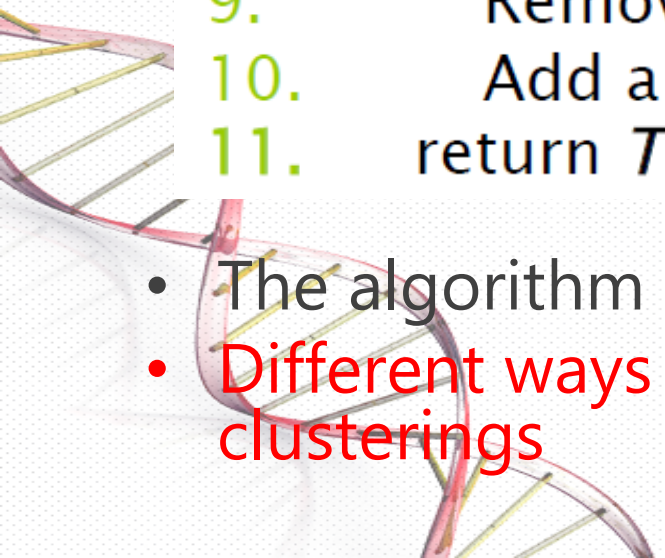
Hierarchical clustering





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 corrsponding to the new cluster C
11. return T



- The algorithm takes a nxn pairwise distance matrix d as an input.
- Different ways to define distances between clusters may lead to different clusterings

Hierarchical clustering: Recomputing distances

$$d_{\min}(C, C^*) = \min d(x, y)$$

for all elements x in C and y in C*

- Distance between two clusters is the **smallest** distance between any pair of their elements (**single-linkage**)

$$d_{\max}(C, C^*) = \max d(x, y)$$

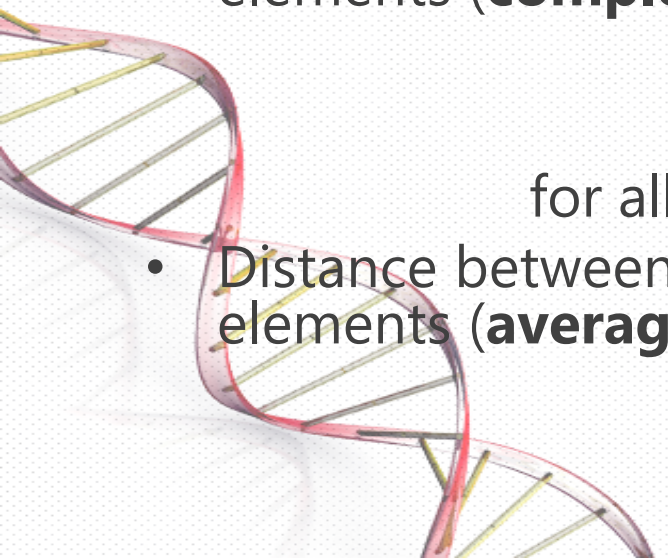
for all elements x in C and y in C*

- Distance between two clusters is the **largest** distance between any pair of their elements (**complete-linkage**)

$$d_{avg}(C, C^*) = \frac{\sum d(x, y)}{|C||C^*|}$$

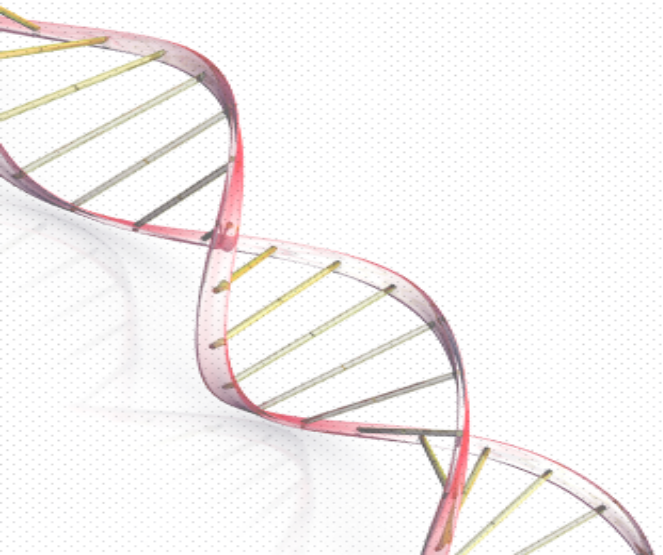
for all elements x in C and y in C*

- Distance between two clusters is the **average** distance between all pairs of their elements (**average-linkage**)



Hierarchical clustering example

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00



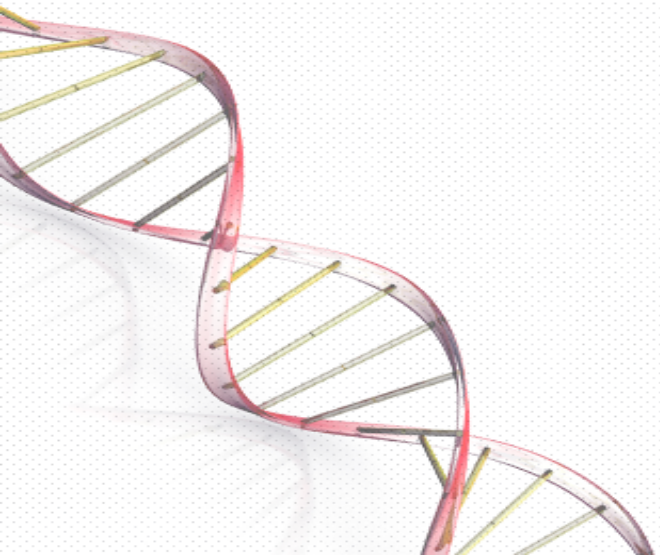
Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Hierarchical clustering example

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	**3.20**	**2.50**	**2.24**	0.00	1.00
E	4.24	3.54	1.41	**1.00**	0.00

Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0



Hierarchical clustering example

Min Distance (Single Linkage)

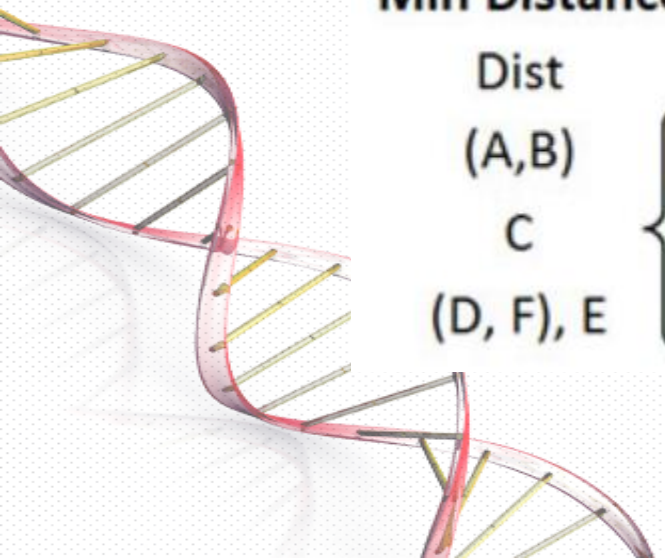
Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	**4.95**	0	2.24	1.41
(D, F)	**2.50**	2.24	0	1.00
E	**3.54**	1.41	1.00	0

Min Distance (Single Linkage)

Dist	(A,B)	(D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	**2.50**	0.00

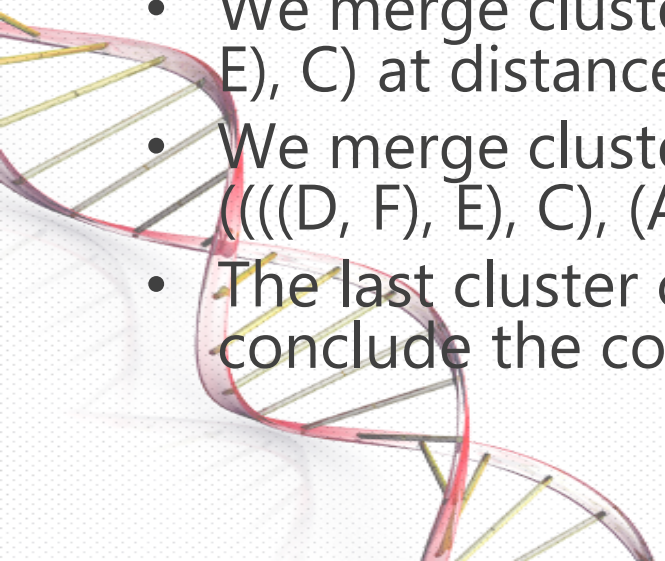
Min Distance (Single Linkage)

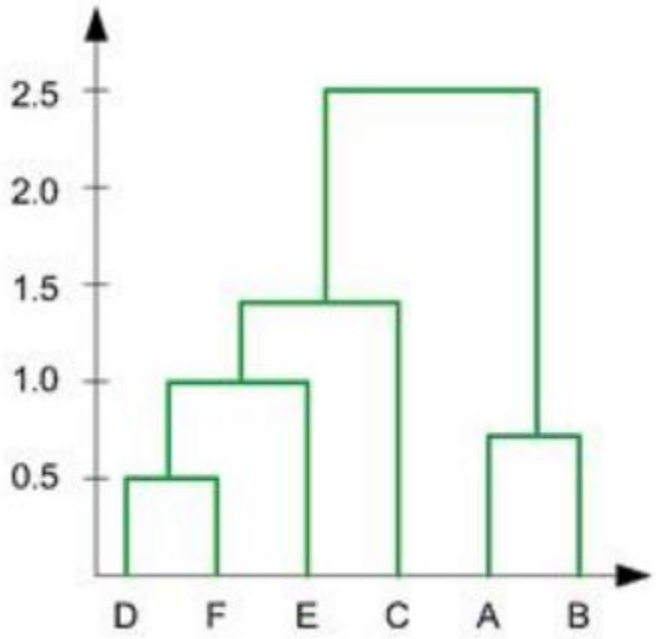
Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	**2.50**	**1.41**	0.00



Hierarchical clustering example

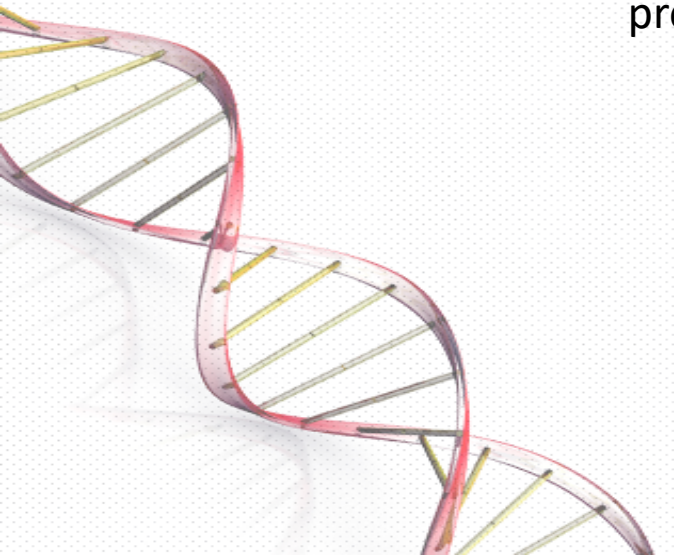
- In the beginning we have 6 clusters: A, B, C, D, E and F
- We merge cluster D and F into cluster (D, F) at distance 0.5
- We merge cluster A and cluster B into (A, B) at distance 0.71
- We merge cluster E and (D, F) into ((D, F), E) at distance 1.0
- We merge cluster ((D, F), E) and C into (((D, F), E), C) at distance 1.41
- We merge cluster (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance 2.5
- The last cluster contain all the objects, thus conclude the computation

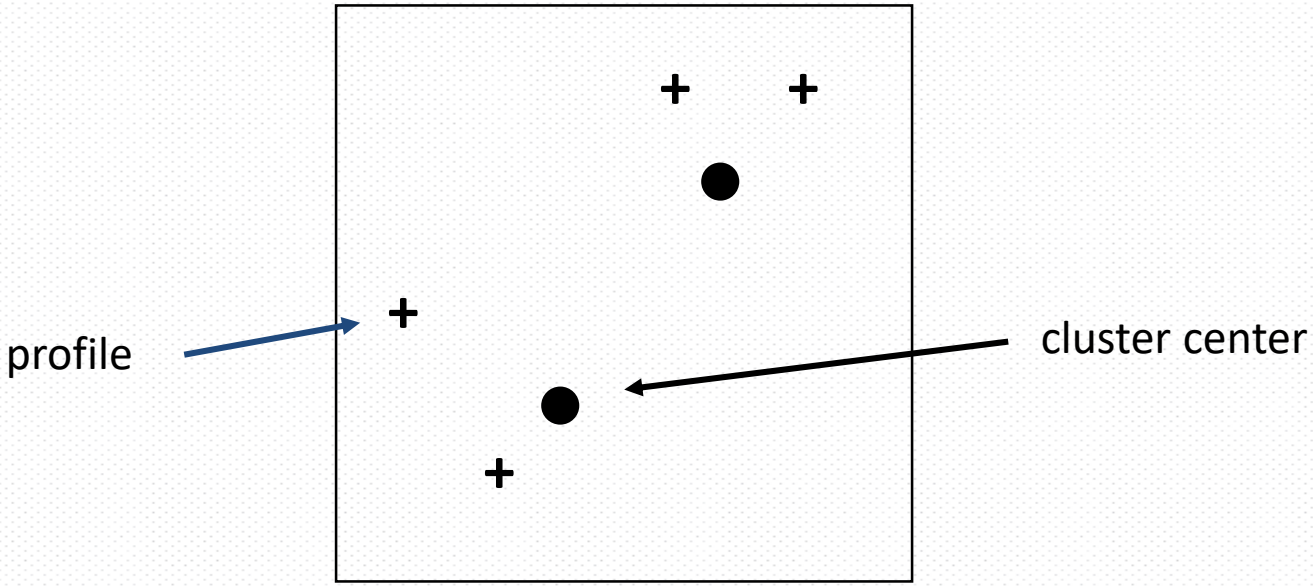




K-Means clustering

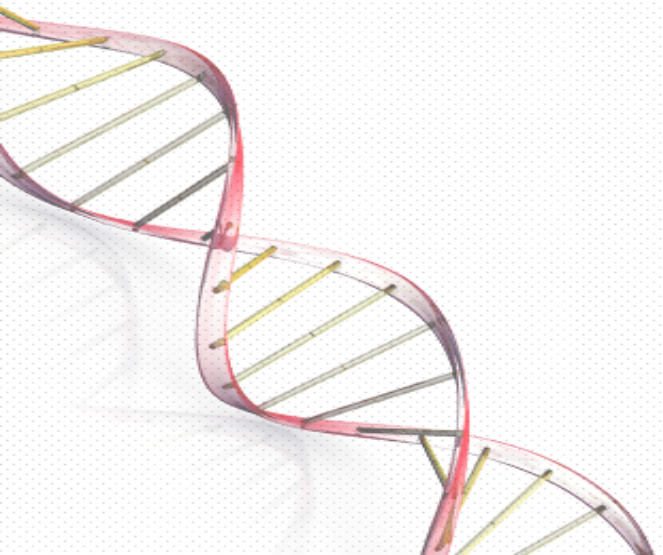
- consider an example in which our vectors have 2 dimensions

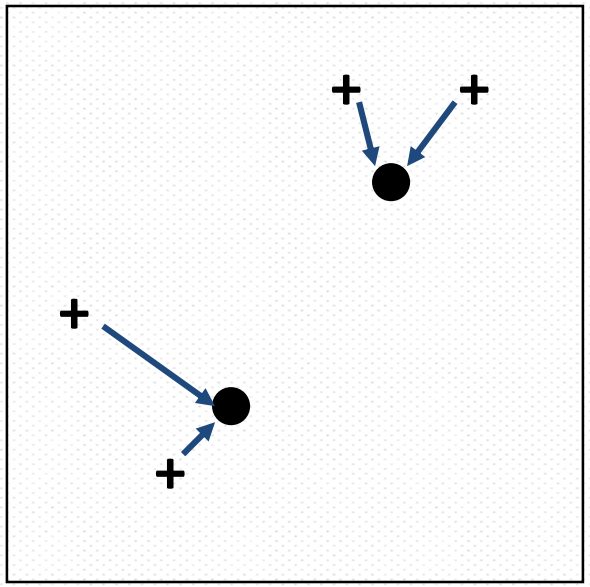




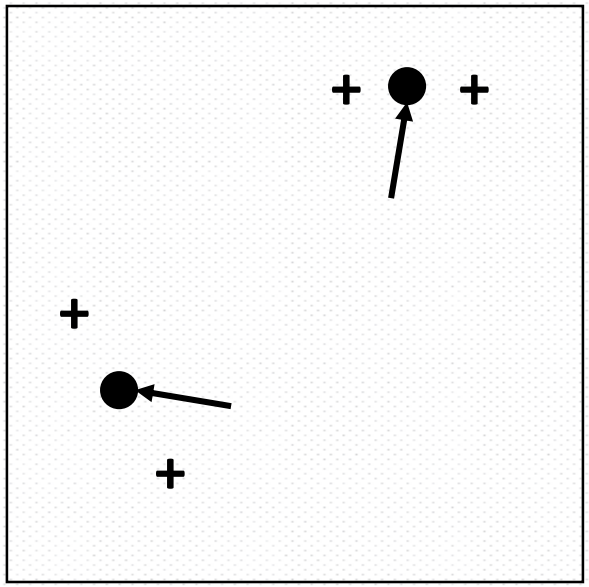
K-Means clustering

- each iteration involves two steps
  - assignment of profiles to clusters
  - re-computation of the cluster centers (means)





assignment

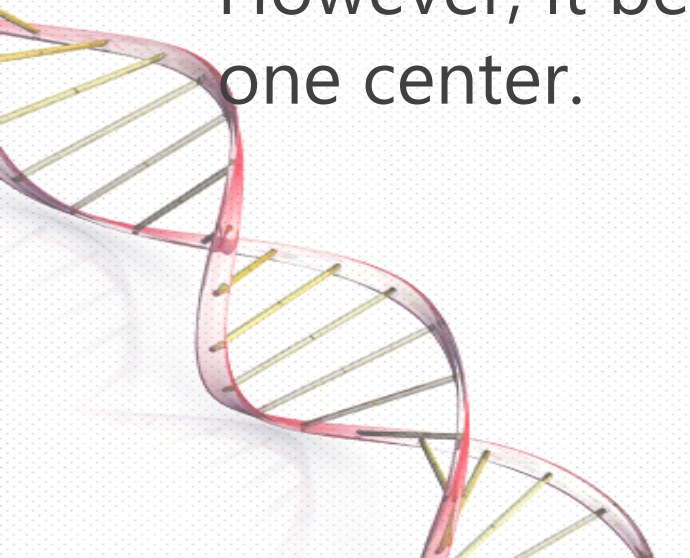


re-computation of cluster centers

1-Means clustering problem

- **Input**: A set, ***V***, consisting of *n* points
- **Output**: A single point ***x*** (cluster center) that minimizes the squared error distortion *d(**V**,**x**)* over all possible choices of ***x***

- 1-Means Clustering problem is easy.
- However, it becomes very difficult (NP-complete) for more than one center.



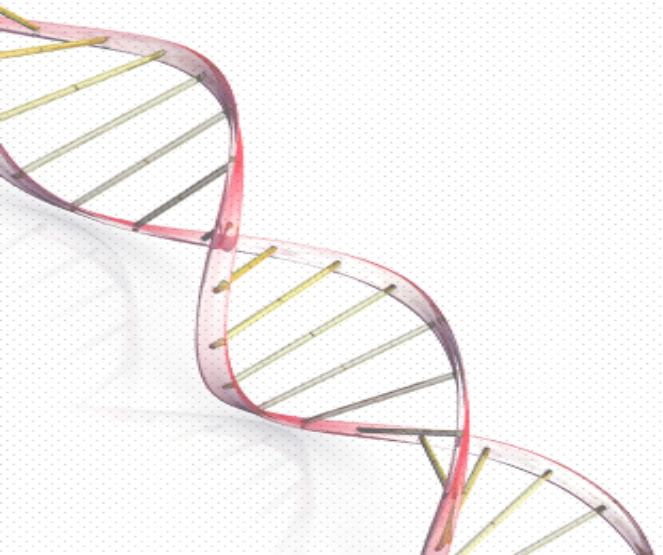
K-Means clustering

- Dissimilarity measure is the Euclidean distance
- Minimizes *within-cluster scatter* defined as

$$\sum_{k=1}^{K} \sum_{i:C(i)=k} \|x_i - f_k\|^2$$

Center of cluster k

Sum over all points assigned to cluster k

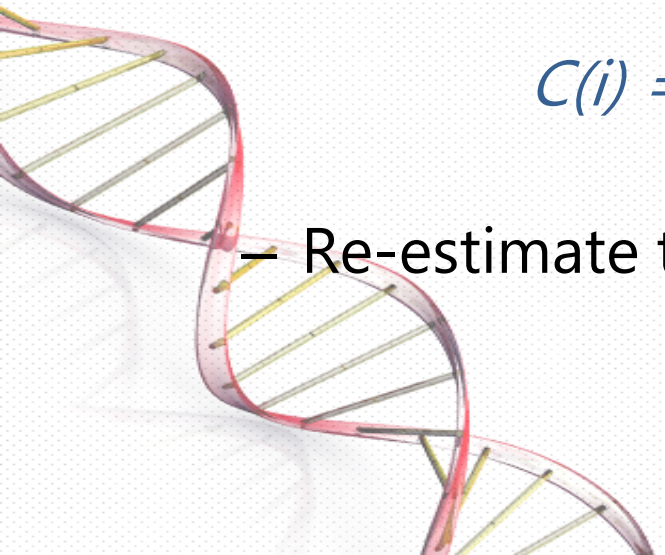


K-Means clustering algorithm

- Input: K, number of clusters, a set $X=\{x_1,.. x_N\}$ of data points, where x_i are p-dimensional vectors
- Initialize
  - Select initial cluster means $f_1, \cdots.., f_K$
- Repeat until convergence
  - Assign each x_i to cluster $C(i)$ such that

$$C(i) = \text{argmin}_{1\le k\le K} \| x_i - f_k \|^2$$

  - Re-estimate the mean of each cluster based on new members



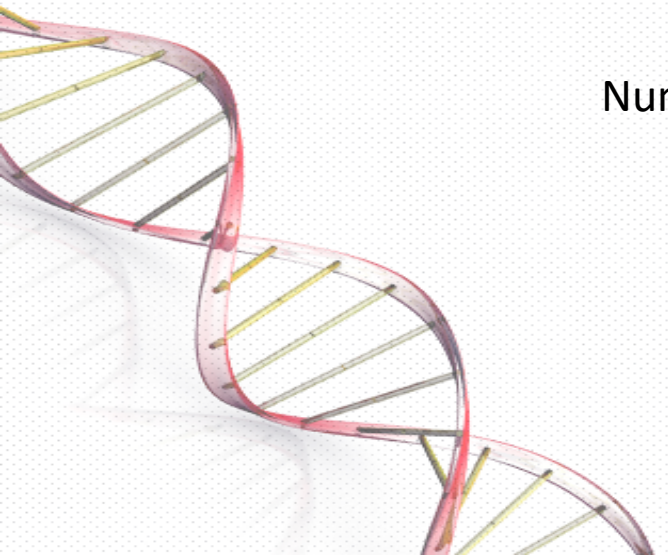
K-means: updating the mean

- To compute the mean of the k^{th} cluster

$$f_k = \frac{1}{N_k} \sum_{i:C(i)=k} x_i$$

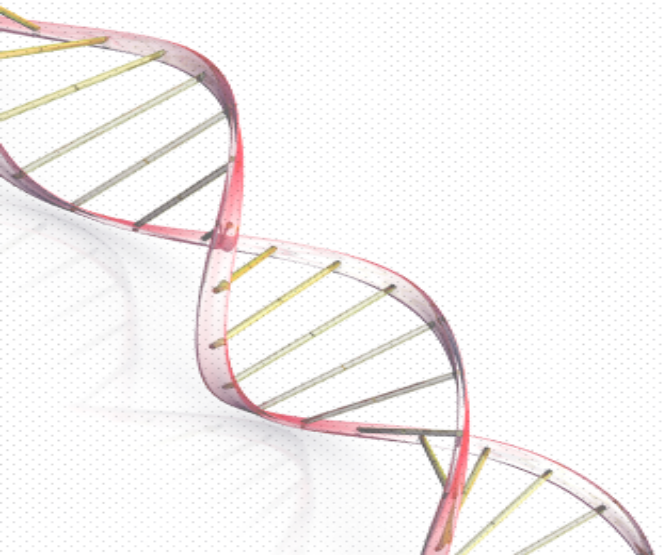
Number of points in cluster k

All points in cluster k



K-means stopping criteria

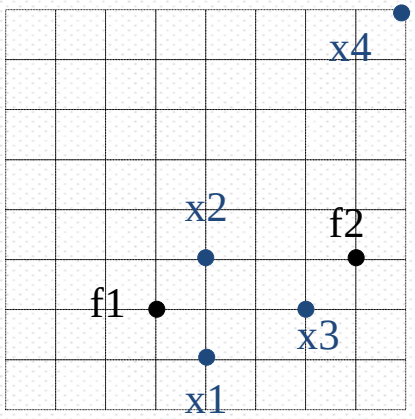
- Assignment of objects to clusters don’t change
- Fix the max number of iterations
- Optimization criterion changes by a small value



K-means Clustering Example

- Given the following 4 instances and 2 clusters initialized as shown.

$$\text{dist}(x_i, x_j)^2 = \|x_i - x_j\|^2$$

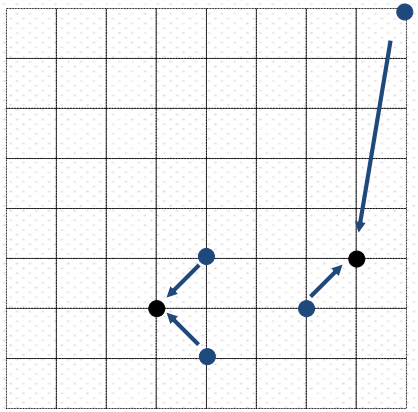


$dist(x_1, f_1)^2 = 2, \quad dist(x_1, f_2)^2 = 13$

$dist(x_2, f_1)^2 = 2, \quad dist(x_2, f_2)^2 = 9$

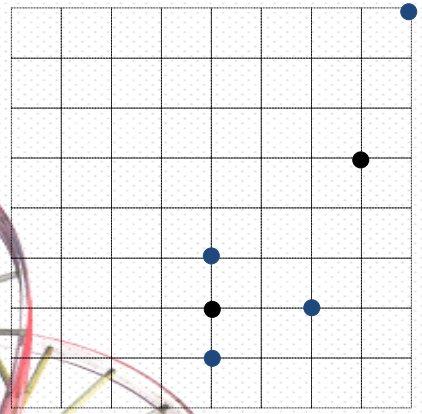
$dist(x_3, f_1)^2 = 9, \quad dist(x_3, f_2)^2 = 2$

$dist(x_4, f_1)^2 = 41, \quad dist(x_4, f_2)^2 = 16$



$$f_1 = \left(\frac{4+4}{2}, \frac{1+3}{2}\right) = (4,2)$$

$$f_2 = \left(\frac{6+8}{2}, \frac{2+8}{2}\right) = (7,5)$$

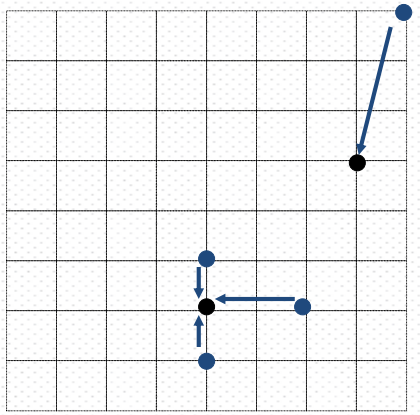


$dist(x_1, f_1)^2 = 1, \quad dist(x_1, f_2)^2 = 25$

$dist(x_2, f_1)^2 = 1, \quad dist(x_2, f_2)^2 = 13$

$dist(x_3, f_1)^2 = 4, \quad dist(x_3, f_2)^2 = 10$

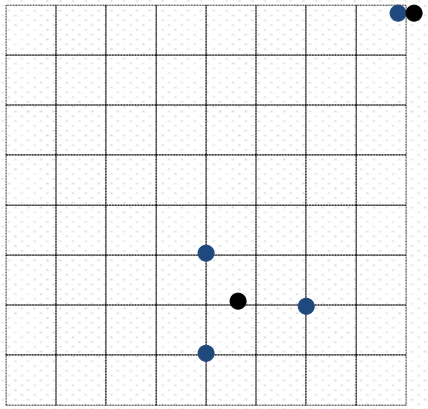
$dist(x_4, f_1)^2 = 52, \quad dist(x_4, f_2)^2 = 10$



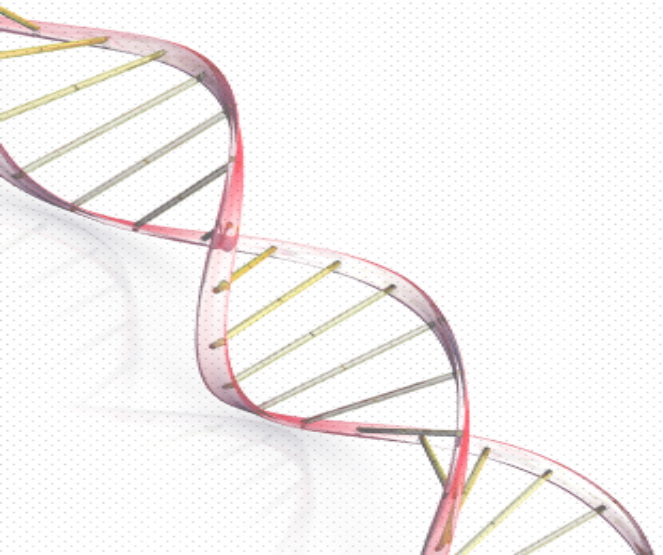
$$f_1 = \left(\frac{4+4+6}{3}, \frac{1+3+2}{3}\right) = (4.67,2)$$

$$f_2 = \left(\frac{8}{1}, \frac{8}{1}\right) = (8,8)$$

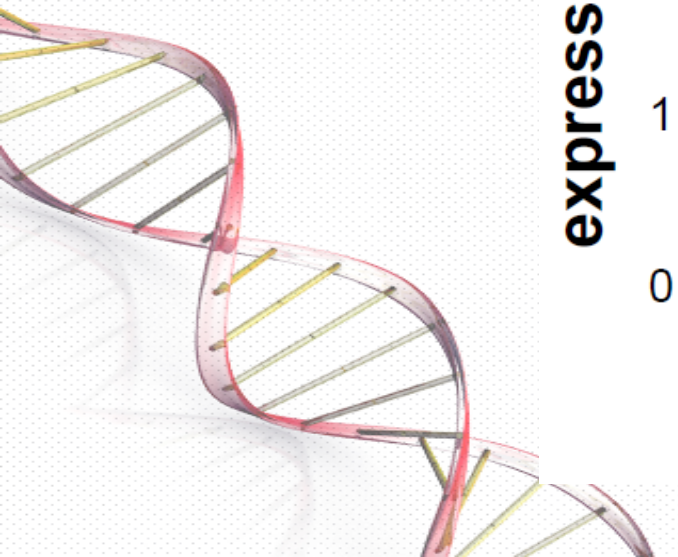
K-means Clustering Example

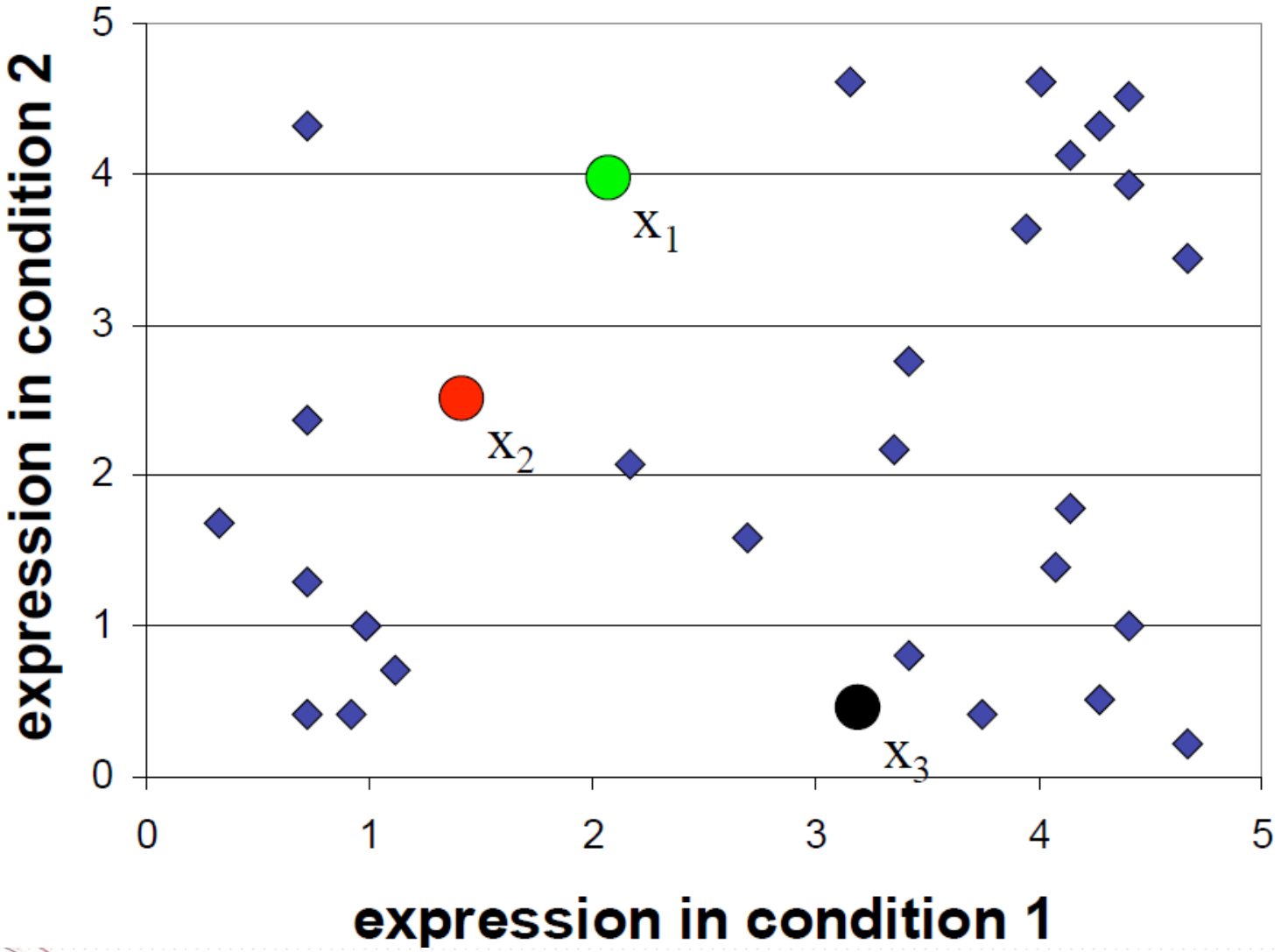


assignments remain the same,
so the procedure has converged

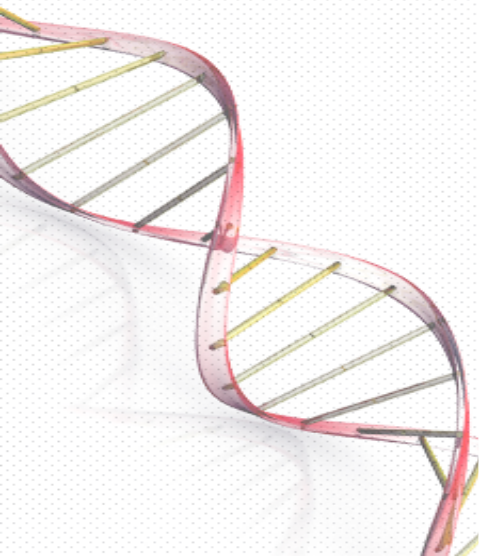


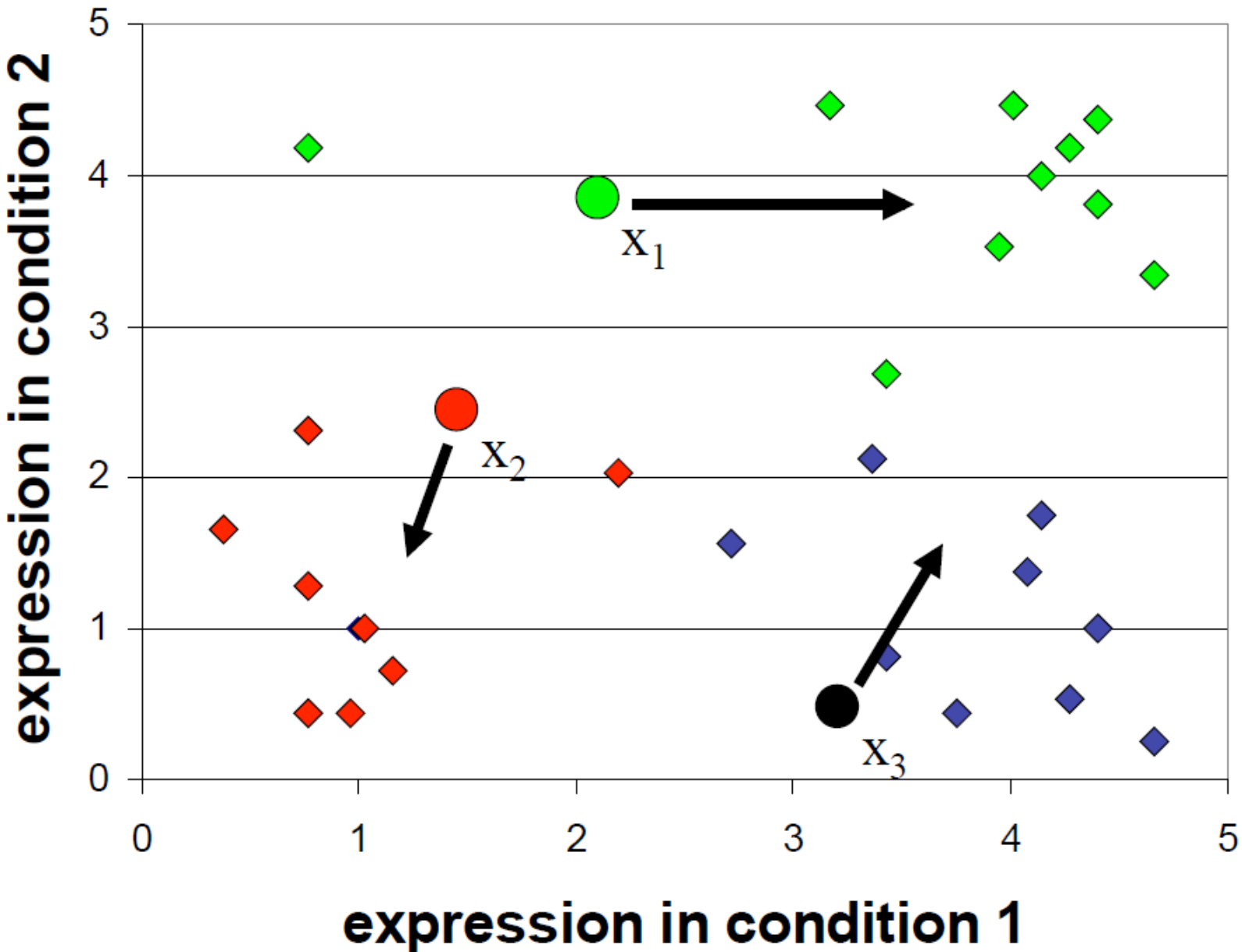
K-means Gene Expression Example



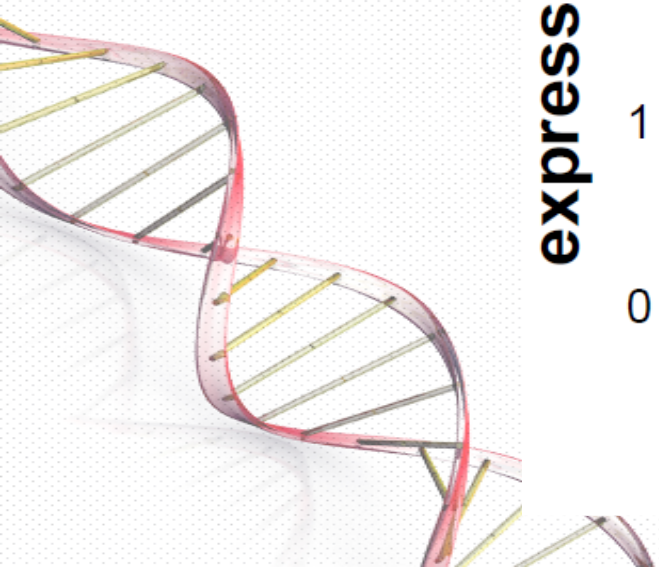


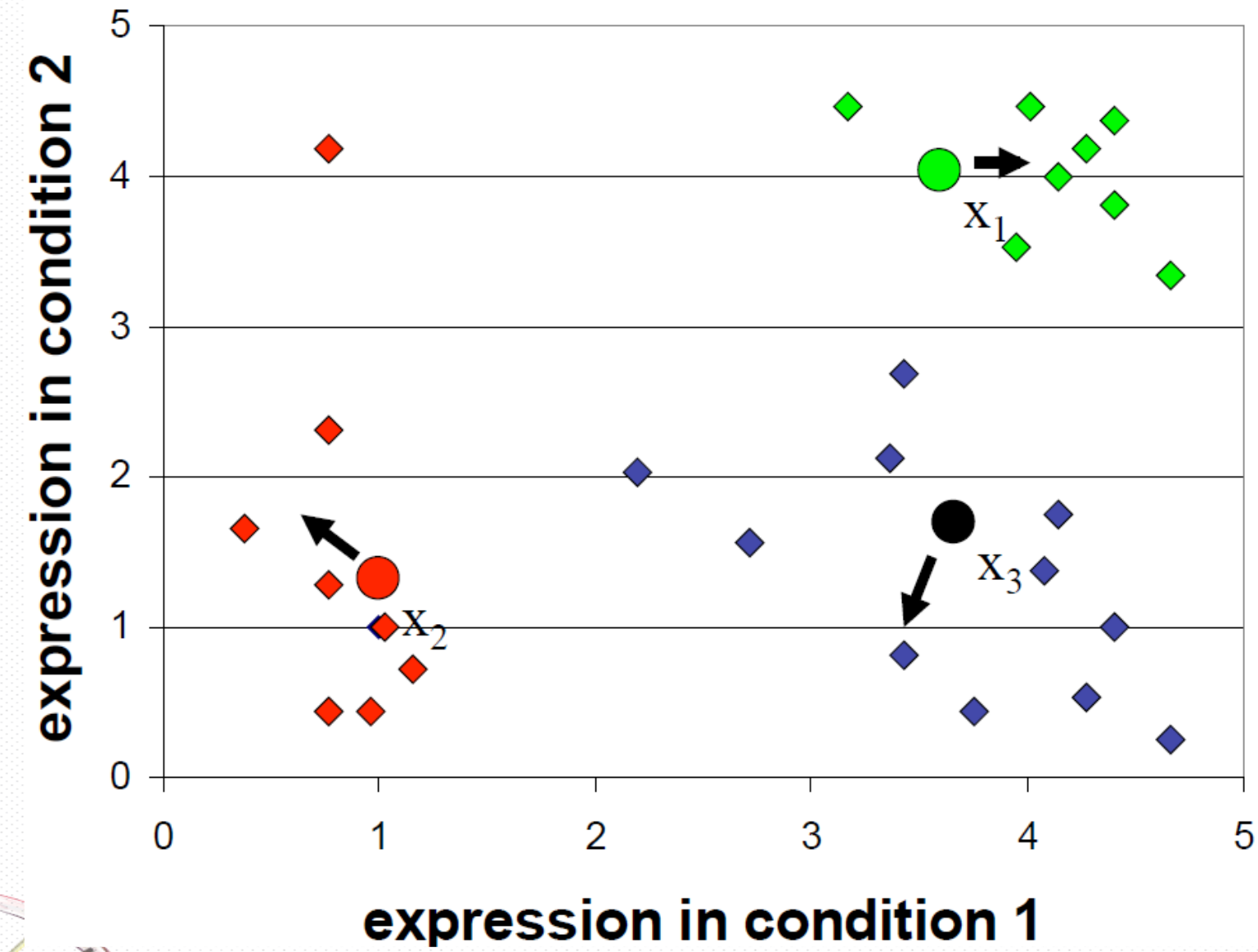
K-means Gene Expression Example



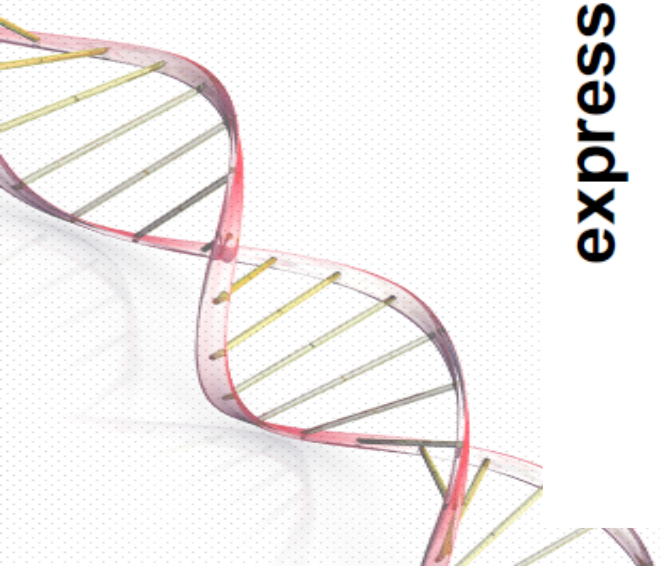


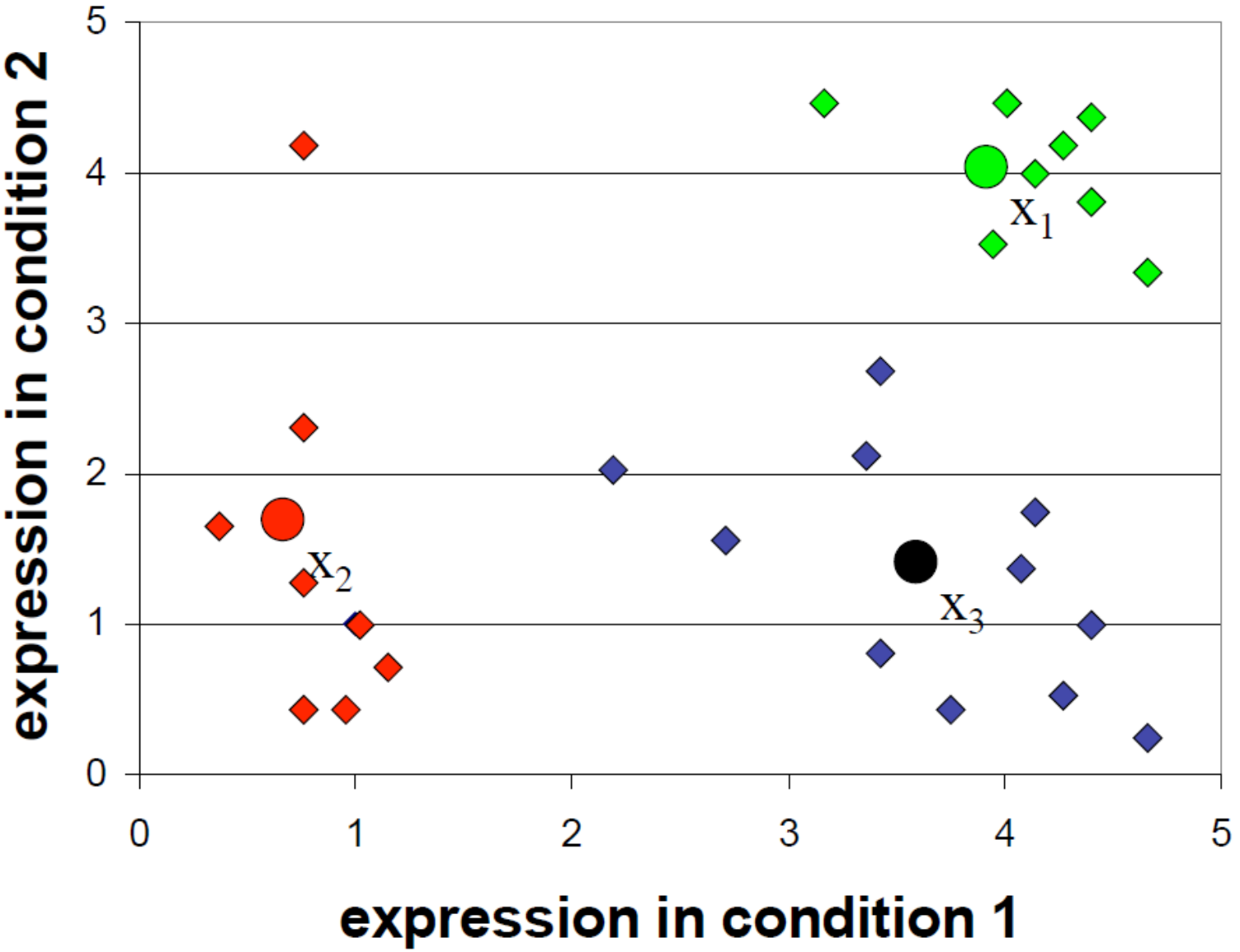
K-means Gene Expression Example





K-means Gene Expression Example





Clustering vs classification

- In the terminology of machine learning
  - clustering is unsupervised learning
  - classification is supervised learning
- Examples
  - Clustering protein sequences to define families
  - Classification: assign a new protein sequence to one of the PFAM families.