COSC 348: Computing for Bioinformatics

Lecture 18:
Microarray Data Analysis: clustering

Lubica Benuskova

http://www.cs.otago.ac.nz/cosc348/

Matrix of selected genes

- We can treat rows/columns of numbers as *n*-dimensional vectors.

Expression profiles (vectors) of genes in condition 1

Expression profiles (vectors) of genes in condition 2

Set of *N* selected genes

ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	436.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	361.9	354.6	339.1	180.6	432.7	210.2	83.6
98590_at	1,066	746.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	585.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434_at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5

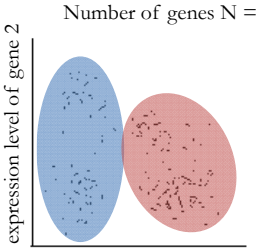
Gene expression profiles

- Each gene has its own **expression profile**, i.e. an array of values for each condition.
- The idea behind clustering is that genes with similar profiles can be grouped together.
- This can help to identify the function of "mystery" genes:
  - if a gene with an unknown function has a similar expression profile to a collection genes, then there is a good chance that the "mystery" gene also plays a role in the studied process.
- It is assumed that the genes with similar expression profiles are co-expressed, e.g. their expression may co-regulated by the same transcription factors.

Cluster analysis: definition

- Clustering is an exploratory procedure which provides a method for grouping objects based on some measure of similarity.
- The goal is to find groups of points that are close to each other and represent a class of objects, e.g. cluster of all subjects which are healthy versus a cluster of subjects who have genetic disease.

Each point (vector) represents 1 subject. Number of genes N = 2.



Cluster analysis: assumptions

- No assumptions are made about the number of groups/clusters, or the group/cluster structure.
- Grouping is done on the basis of similarities, or dissimilarities, i.e. distances between vectors (points in n-dim space).
- A quantitative scale (metric) is used to measure the closeness or similarity between objects.
- In the microarray setting, the objects (vectors) are the expression profiles of the genes in the experiment.

[Mik Black, bioc354 lecture]

Cluster analysis: distance measure

- One popular measure of the closeness between two vectors is the *Euclidean distance*:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

- Another popular measure is the *Pearson's correlation coefficient* (σ is the standard deviation):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

- There are other distance (correlation) measures. The choice depends on the measurement scale, heuristic knowledge, or we can experiment with different measures to see, which one yields better results. Note that the metrics are symmetric.

K-means clustering

- *K*-means clustering is a method of cluster analysis, which aims to partition *N* vectors into *K* clusters, in which each vector belongs to the cluster with the nearest mean (i.e. center of a cluster).
- It attempts to find the centers of natural clusters in the data by the iterative refinement.
- It is the most popular "bottom-up" clustering algorithm.
- If *K* is the number of clusters, *n* is the dimensionality of vectors, and *N* the number of vectors, the problem can be exactly solved in time $O(N^{nK+1} \log N)$.

K-means clustering

- Vector clustering for *K* centers:
  - *K*-means clustering partitions the vectors into *K* clusters $\mathbf{C}=\{C_1,C_2,\ldots,C_K\}$ so as to minimize the sum of distances between all vectors belonging to the cluster and the centre of the cluster

$$\underset{\mathbf{C}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\mathbf{x}_p \in C_k} d(\mathbf{x}_p, \mathbf{c}_k)$$

  - Cluster C_k has the centre $\mathbf{c}_k$
  - Let us have *P* vectors in total (*p* is the index of input vector)
  - But how do we know which vectors belong to the cluster *k*?

K-means clustering algorithm

- Randomly choose *K* input vectors from the data, which will be the starting centres of clusters.
- Then for each input vector $\mathbf{x}_p$ do:
  - Find the centre $\mathbf{c}_k$, which is closest to $\mathbf{x}_p$
  - Create a new centre $\mathbf{c}_k$ by changing each coordinate *i*, so that:

$$c_{ki} \leftarrow c_{ki} + \alpha\left(x_{pi} - c_{ki}\right)$$

  - Stopping criterion: either we can make *a* linearly decreasing to zero or we can stop when centres do not move significantly. In general, we choose initial $\alpha \in (0, 1]$.

K-means clustering

- *K*-means clustering is run for different number of clusters *K* as we do not know ahead how many natural clusters are in the data.
- There is no guarantee it will converge to the global optimum.
- The result depends on the initial random choice of centers of clusters.
- As the algorithm is usually very fast, it is common to run it multiple times with different starting centers.
- Particular value of $\alpha \in (0, 1]$ must be also found by experimentation.

K-means++ algorithm

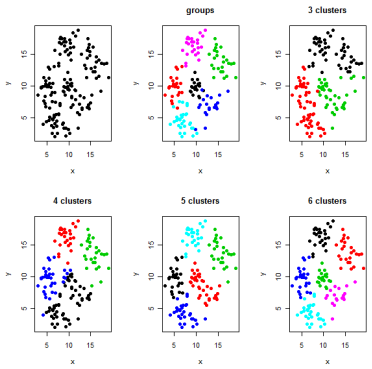
- *K*-means++ is designed to improve the initial choice of cluster centers based on spreading the *K* initial cluster centers away from each other:

1. Choose one center uniformly at random from among the data points.
2. For each data point x_p, compute $d(x_p, c_k)$, the distance between x_p and the nearest center that has already been chosen.
3. Add one new data point at random as a new center, using a weighted probability distribution, where a new point x_p^* is chosen with probability proportional to the distance $d(x_p^*, c_k)$.
4. Repeat Steps 2 and 3 until *K* centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard *K*-means clustering.

K-means++ clustering

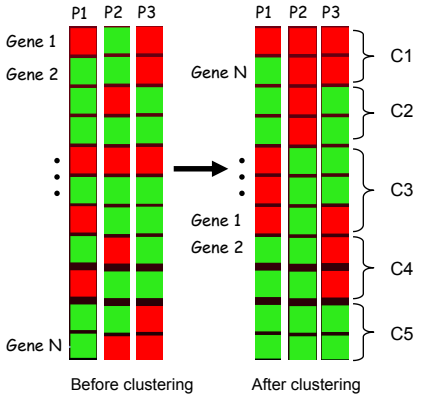
- This seeding method gives out considerable improvements in the final error of *K*-means.
- Although the initial selection in the algorithm takes extra time, the *K*-means part itself converges very fast after this seeding and thus the algorithm actually lowers the computation time too.
- The authors tested their method with real and synthetic datasets and obtained typically a 2-fold improvements in speed, and for certain datasets close to 1000-fold improvements in error.
- It was proposed in 2007 by David Arthur and Sergei Vassilvitskii as a way of avoiding the sometimes poor clusterings found by the standard *K*-means algorithm.

How to choose *K* ?



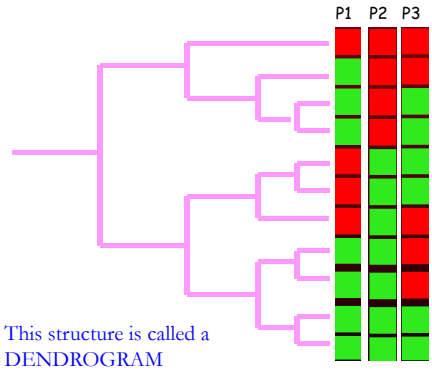
- How many clusters are there? We do not know.
- We have to apply some heuristics to choose *K* (e.g. *K* = number of conditions) or experiment with different *K*s.

K-means clustering of gene profiles



- Let us have spotted microarrays. Red means high expression, green means low expression.
- We have clustered genes into 5 clusters according to similarity of their profiles across patients P1, P2 and P3.

Hierarchical clustering based on *K*-means

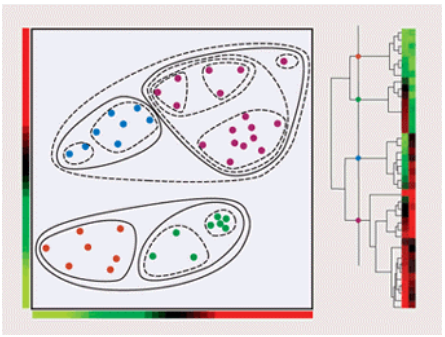


This structure is called a DENDROGRAM

- First we perform the *K*-means clustering
- Then we hierarchically arrange the clusters according to linkage criteria
- Result: a tree of distances between clusters of vectors

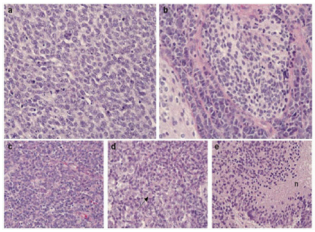
Hierarchical clustering

- Another illustration of hierarchical organisation of clusters produced by *K*-means first (*K* = 9).



Brain tumour data

- Brain data, Scott Pomeroy et al, Nature (vol. 415), Jan 2002
  - 99 samples, about 7000 genes, 5 classes of brain tumour
  - Authors have selected 50 genes that are most differentially expressed based on *t*-test.



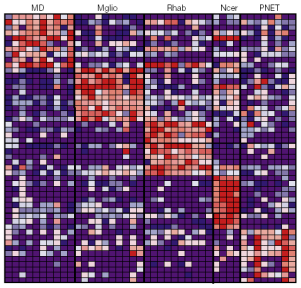
Photomicrographs of tumours:

a, MD (medulloblastoma) classis
b, MD desmoplastic
c, PNET
d, rhabdoid tumour
e, glioblastoma
Analysis also used Normal cerebella tissue (Ncer), not shown here

Real example of *K*-means clustering

- They performed gene clustering on selected genes for *K* = 5.
- As a result, each condition (disease) may involve different cluster of genes, which are differentially expressed.
- **The goal** of gene clustering: to reveal the clusters of differentially expressed genes, which characterize each condition.

[Pomeroy et al., Nature 415: 436-442, 2002]



Rows: selected genes (profile vectors)
Columns: patients
Classes: 4 brain tumour types + Ncer