# Clustering Analysis
# Babelomics 5.0

Cankut ÇUBUK

March 3rd, 2016


GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

# Outline

- Introduction
- Types of clustering*
- Methods*
  - UPGMA
  - SOTA
  - K-Means
- Paramaters*
  - Distance
  - K-value

*Babelomics 5.0 based

# What is clustering analysis?

**Cluster** is a group of similar things which have a relatively close association.

**Clustering analysis** is grouping a set of data objects into clusters.

- In our case genes and/or samples

| | sampleA | sampleB | sampleC | sampleD | sampleE |
|---|---|---|---|---|---|
| gene1 | 47 | 20 | 24 | 36 | 12 |
| gene2 | 35 | 47 | 33 | 47 | 42 |
| gene3 | 39 | 19 | 21 | 18 | 46 |
| gene4 | 38 | 12 | 44 | 16 | 22 |
| gene5 | 19 | 14 | 16 | 20 | 31 |
| gene6 | 19 | 26 | 36 | 18 | 12 |
| gene7 | 24 | 38 | 46 | 14 | 24 |

# What is clustering analysis?

Clustering is **unsupervised** classification.

- No predefined class.

| | sampleA | sampleB | sampleC | sampleD | sampleE |
|---|---|---|---|---|---|
| gene1 | 47 | 20 | 24 | 36 | 12 |
| gene2 | 35 | 47 | 33 | 47 | 42 |
| gene3 | 39 | 19 | 21 | 18 | 46 |
| gene4 | 38 | 12 | 44 | 16 | 22 |
| gene5 | 19 | 14 | 16 | 20 | 31 |
| gene6 | 19 | 26 | 36 | 18 | 12 |
| gene7 | 24 | 38 | 46 | 14 | 24 |

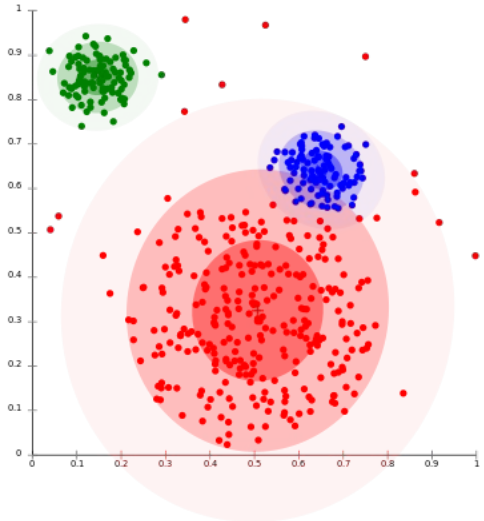
| | sample_name | sample_type |
|---|---|---|
| 1 | "sampleA" | "Tumor" |
| 2 | "sampleB" | "Normal" |
| 3 | "sampleC" | "Tumor" |
| 4 | "sampleD" | "Normal" |
| 5 | "sampleE" | "Tumor" |

# What is clustering analysis?

A good clustering method will produce **high quality clusters** with

- High intra-class similarity (Green, Blue >> Red)
- Low inter-class similarity  (Green vs Blue, Green vs Red >> Blue vs Red)

! Depends on your data.



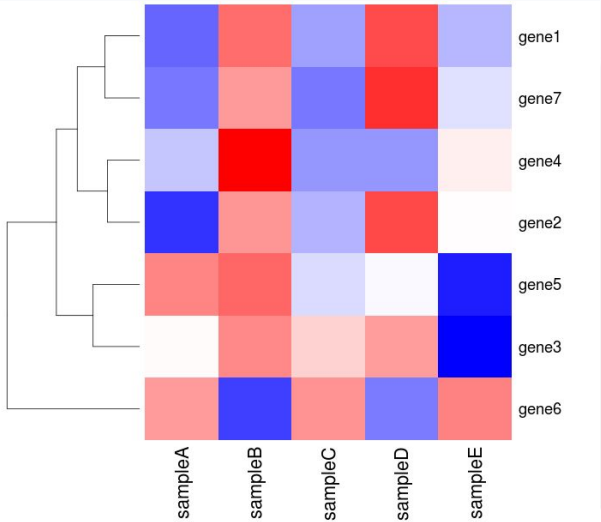
# Which questions can be answered with clustering?

Are there some genes with a similar pattern of gene expression across arrays?

- The unit of analysis is the gene
- Find genes that behave the same across patients
- Indicate possible gene funtionality
- Find temporal patterns of gene expression

Select type of clustering: samples and/or genes

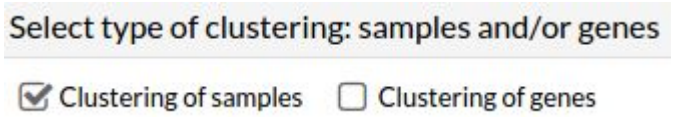
☐ Clustering of samples ☑ Clustering of genes

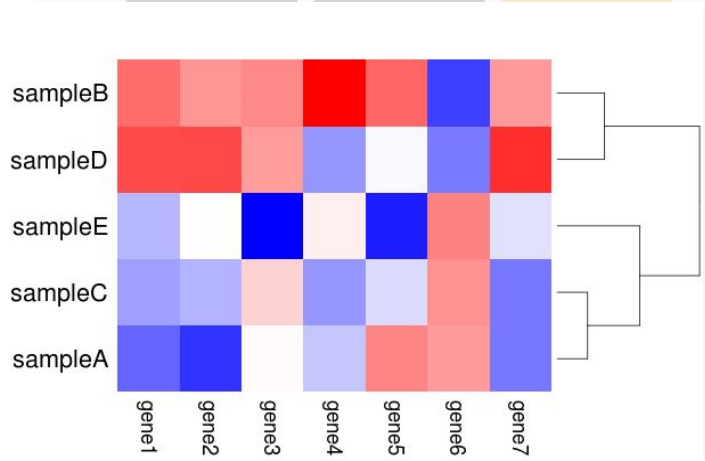


# Which questions can we answer with clustering?

Are there some biological samples with the same pattern of gene expression across genes?

- The unit of analysis is the sample
- Discover new subgroups in a set of patients of the same disease
- Descriptive analysis
- Perform quality control checking
  - Outlier detection
  - Batch effect assessment



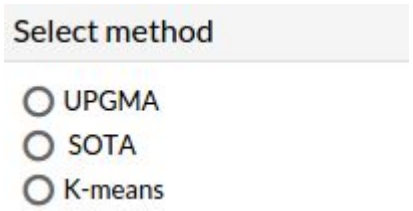



# Clustering Methods in Babelomics

Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

Self-Organizing Tree Algorithm (SOTA)

K-Means

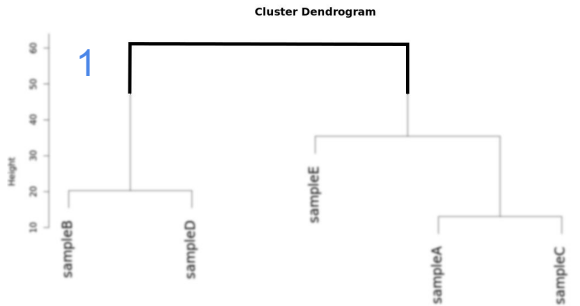


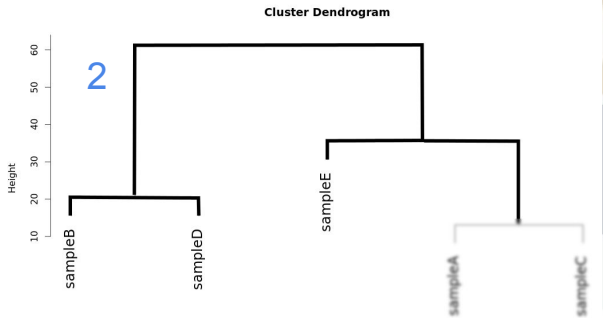
# UPGMA

- UPGMA is a simple agglomerative (bottom-up) hierarchical clustering method.
- This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- It is not the more accurate among the methods but is really extensively used especially for gene expression data. Provides a tree.

# SOTA

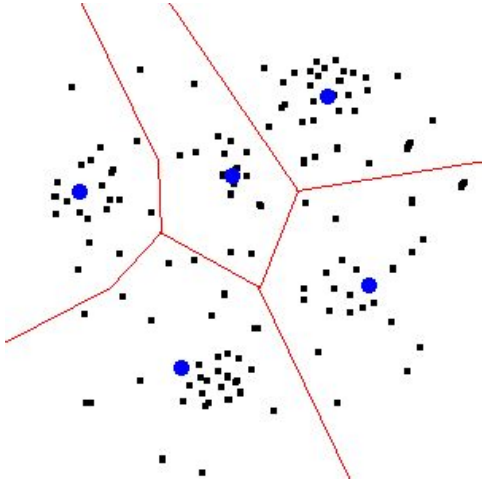
- SOTA starts the classification with a binary topology composed of a root node with two leaves.
- A divisive(top down) method.
- The self-organizing process splits the data (e.g. samples) into two clusters.
- After reaching convergence at this level, the network is inspected.
- If the level of variability in one, or more, terminal nodes is over a given threshold, then, the tree grows by expanding these terminal nodes.
- Provides a tree.

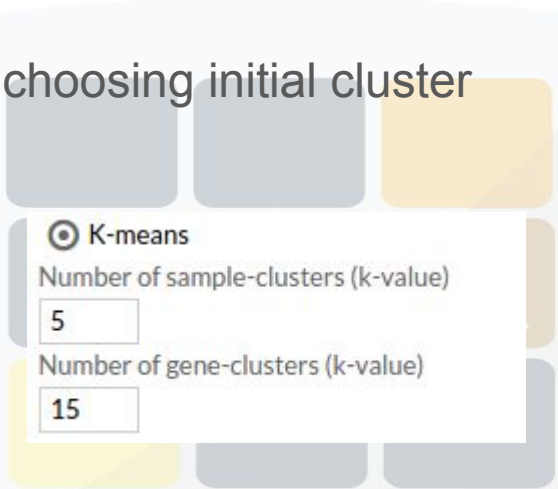


# K-Means

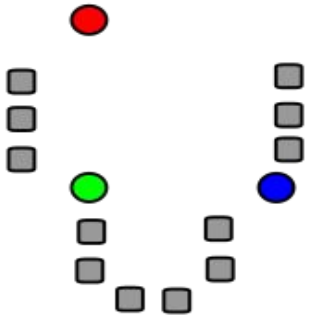
- K-means aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.
- Do not provide a tree.
- Usually need the number of cluster to be set.
- Its result is very sensitive to the initialization step: choosing initial cluster centers.





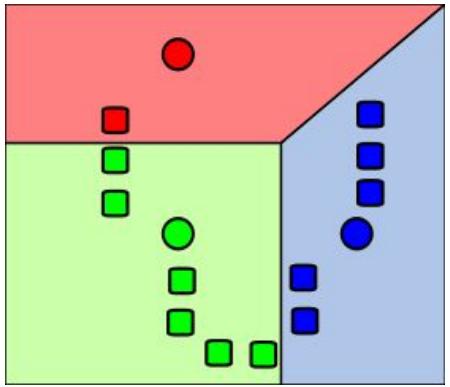
# K-Means

1



Red,Blue,Green circles are initial cluster centers.

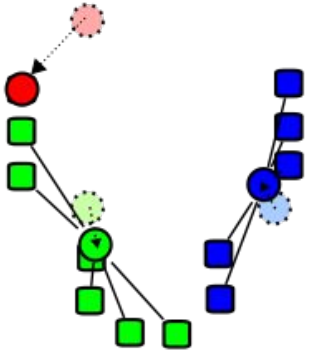
2



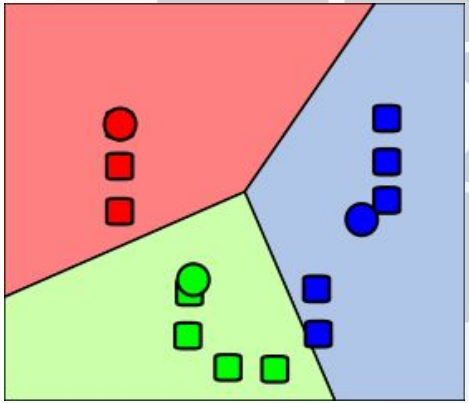
We find closest observations to initial centers.

We calculate center of new clusters.

3



4



Repeat step 2 and 3 until no changes occur.

# Distance Parameters

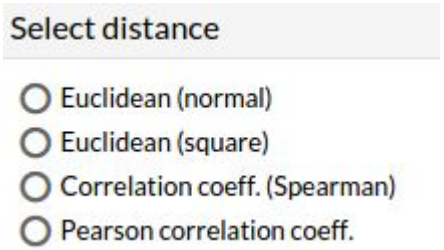
Different distances account for different properties.

1.Euclidean

- Normal
- Squared

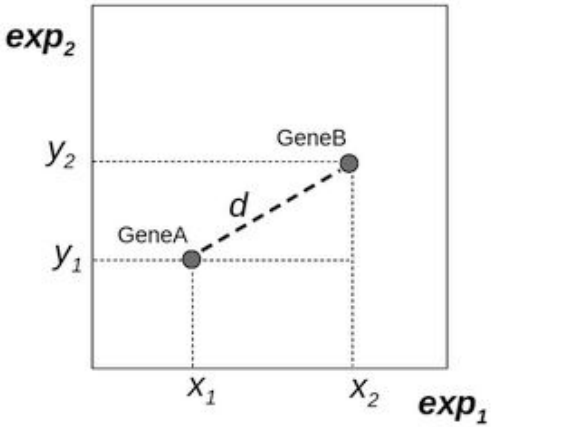
2.Correlation coefficient

- Spearman
- Pearson



# Distance Parameters

## 1.Euclidean




$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

| #Distance Matrix | GeneA | GeneB |
|---|---|---|
| GeneA | 1 | 12.85 |
| GeneB | 12.85 | 1 |

| | GeneA | GeneB | (GeneA-GeneB) Squared |
|---|---|---|---|
| Sample1 | 35 | 30 | 25 |
| Sample2 | 23 | 33 | 100 |
| Sample3 | 47 | 45 | 4 |
| Sample4 | 17 | 23 | 36 |
| Sum | | | 165 |
| Square root of sum | | | 12.85 |

# Distance Parameters

## 2.Correlation coefficient

- **Pearson**

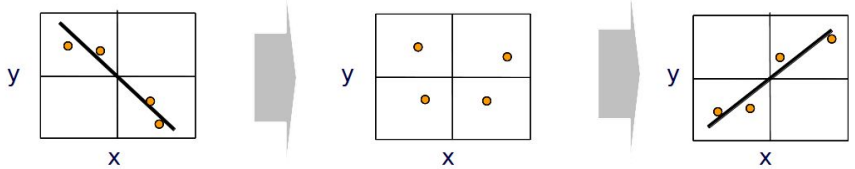
$$r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

| #Distance Matrix | GeneA | GeneB |
|---|---|---|
| GeneA | 1 | (1 - 0.98) |
| GeneB | (1 - 0.98) | 1 |

| | GeneA | GeneB | X squared | y squared | xy |
|---|---|---|---|---|---|
| Sample1 | 35 | 30 | 1225 | 900 | 1050 |
| Sample2 | 23 | 33 | 529 | 1089 | 759 |
| Sample3 | 47 | 45 | 2209 | 2025 | 2115 |
| Sample4 | 17 | 23 | 289 | 529 | 391 |
| Sum | | | 4252 | 4523 | 4315 |

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Cor coef. = 4315 / √(4252*4523) = 0.98



# Distance Parameters

## 2.Correlation coefficient

- **Spearman**

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

| | GeneA | RankA | GeneB | RankB | RankA-RankB (d) | d squared |
|---|---|---|---|---|---|---|
| Sample1 | 35 | 3 | 30 | 2 | 1 | 1 |
| Sample2 | 23 | 2 | 33 | 3 | -1 | 1 |
| Sample3 | 47 | 4 | 45 | 4 | 0 | 0 |
| Sample4 | 17 | 1 | 23 | 1 | 0 | 0 |

| #Distance Matrix | GeneA | GeneB |
|---|---|---|
| GeneA | 1 | (1 - 0.8) |
| GeneB | (1 - 0.8) | 1 |

Cor coef. = 1 - (6*(1+1+0+0))/(4(16-1)) = 1 - (12/60) = 1 - 0.2 = 0.8

# Distance Parameters

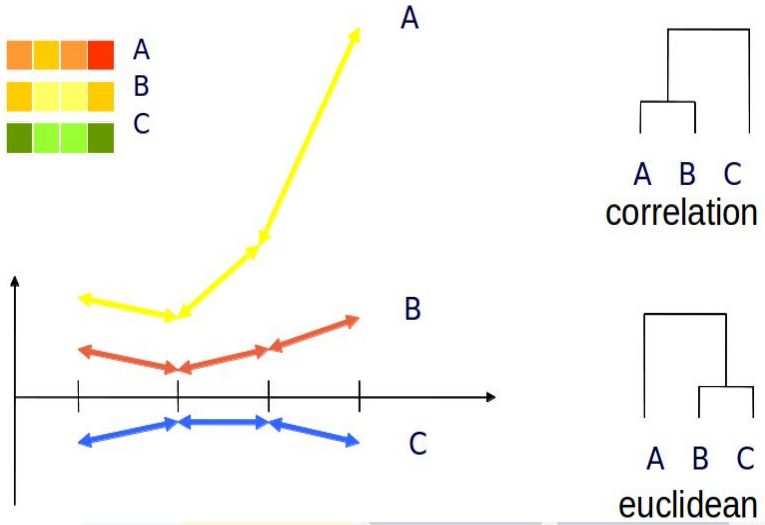
FAQs:

1. Which one is the best?
2. Which one I have to select?
3. Which is the proper question above?

Different definitions of **being close.**

**Correlation**: tendencies

**Euclidean**: global similarity

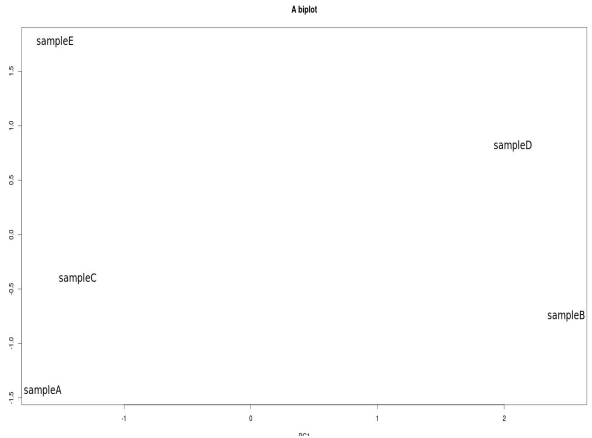


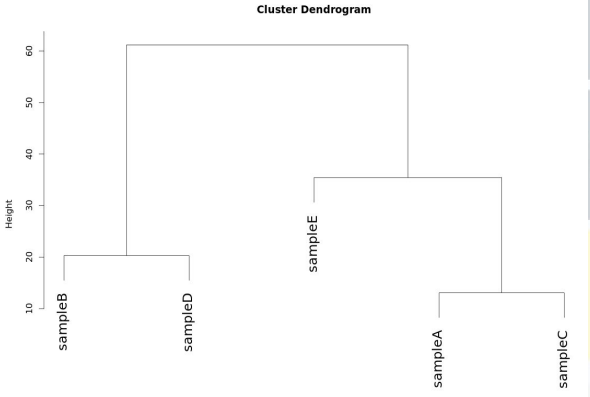
# Visualisation of Clustering Results

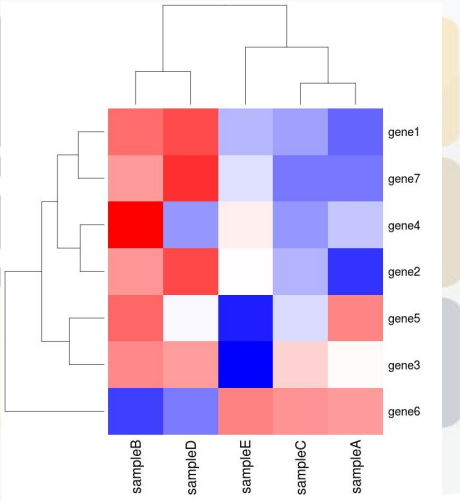
Biplot

Tree Diagram (**Dendrogram**: from Greek dendro "tree" and gramma "drawing")

Heatmap

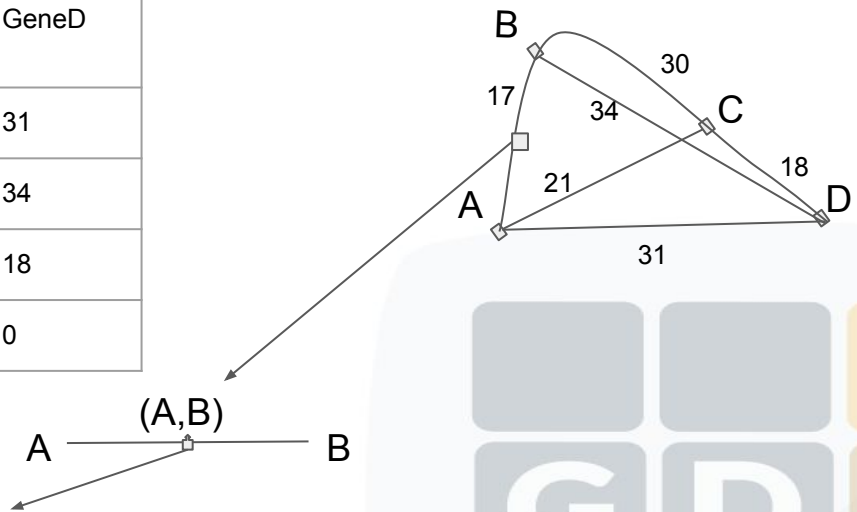


# UPGMA example

| #Distance Matrix | GeneA | GeneB | GeneC | GeneD |
|---|---|---|---|---|
| GeneA | 0 | **17** | 21 | 31 |
| GeneB | **17** | 0 | 30 | 34 |
| GeneC | 21 | 30 | 0 | 18 |
| GeneD | 31 | 34 | 18 | 0 |




D(A,B)=17

**First brach length estimation:** 17/2=8.5

First node where GeneA and GeneB are connected.

Recalculate distance of other genes to this node.

# UPGMA example

*Update the distances*

**D((A,B),C)**=(D(A,C)+D(B,C))/2

=(21+30)/2=25.5

**D((A,B),D)**=(D(A,D)+D(B,D))/2

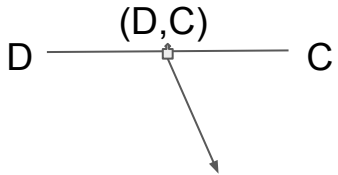
=(31+34)/2=32.5

| **#Distance Matrix** | GeneA | GeneB | GeneC | GeneD |
|---|---|---|---|---|
| GeneA | 0 | **17** | 21 | 31 |
| GeneB | **17** | 0 | 30 | 34 |
| GeneC | 21 | 30 | 0 | 18 |
| GeneD | 31 | 34 | 18 | 0 |

| **#Distance Matrix** | NodeAB | GeneC | GeneD |
|---|---|---|---|
| NodeAB | 0 | 22.5 | 32.5 |
| GeneC | 22.5 | 0 | 18 |
| GeneD | 32.5 | 18 | 0 |

# UPGMA example

**Second brach length estimation:**




D(C,D)=18/2=9.5

*Update the distances*

D((C,D),(A,B))=(D(C,(A,B)) + d(D,(A,B))) / 2

= (22.5 + 32.5) / 2

= 27.5

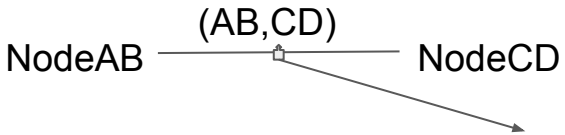
| **#Distance Matrix** | NodeAB | GeneC | GeneD |
|---|---|---|---|
| NodeAB | 0 | 22.5 | 32.5 |
| GeneC | 22.5 | 0 | **18** |
| GeneD | 32.5 | **18** | 0 |

| **#Distance Matrix** | NodeAB | NodeCD |
|---|---|---|
| NodeAB | 0 | 27.5 |
| NodeCD | 27.5 | 0 |

# UPGMA example

**Third brach length estimation:**

**(Our root node)**




D(NodeAB,NodeCD)=27.5 / 2 = 13.75

| **#Distance Matrix** | NodeAB | NodeCD |
|---|---|---|
| NodeAB | 0 | 27.5 |
| NodeCD | 27.5 | 0 |

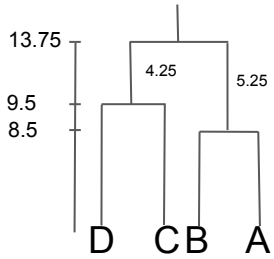
*To create the dendrogram we deduce the other remaining branch lengths:*

NodeRoot-NodeAB = 13.75 - 8.5

= 5.25

NodeRoot-NodeCD = 13.75 - 9.5

= 4.25



# Exercises on Babelomics

http://babelomics.bioinfo.cipf.es/

# Exercises

1) Go to http://bioinfo.cipf.es/gda16/doku.php/program

2) Download GDA16_TCGA_265_mod_gene_BRCA_subtype_HER_Basal_Normal.txt

Data description: RNA-Seq data of 30 Breast Invasive Carcinoma (BRCA) samples taken from The Cancer Genome Atlas (TCGA) data portal.

Contains 10 normal samples, 20 tumor samples with 2 subtypes (Basal-like and Her2-enriched).

3) Upload your file to Babelomics 5.0.

   Go to section Expression>Clustering

4) Cluster samples with given paramaters.

UPGMA + Euclidean (square)

UPGMA + Correlation coeff. (Spearman)

Which distance parameter is better for proper clustering?

# Exercises

5) Repat the analysis using the same distance parameters and SOTA method.

SOTA + Euclidean (square)

SOTA + Correlation coeff. (Spearman)

Do the results change based on the method or the distance parameter?

6) Try to cluster your samples with K-means.

Set k-value 6 and use Correlation coeff. (Spearman)

Repeat the same analysis with k-value 3.

Check the results of K-means.

Are the results acceptable?

Is the dendrogram representing any hierarchy between the samples?

# Exercises

7) Repeat the step 6 with k-value 3.

Did your result same as previous one?

8) ) Try to cluster your samples with K-means.

Set k-value 2 and use Correlation coeff. (Spearman).

Can we say that K-means is good to distinguish tumor from normal?