Clustering Analysis
Babelomics 5.0

Cankut ÇUBUK
March 9th, 2017

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

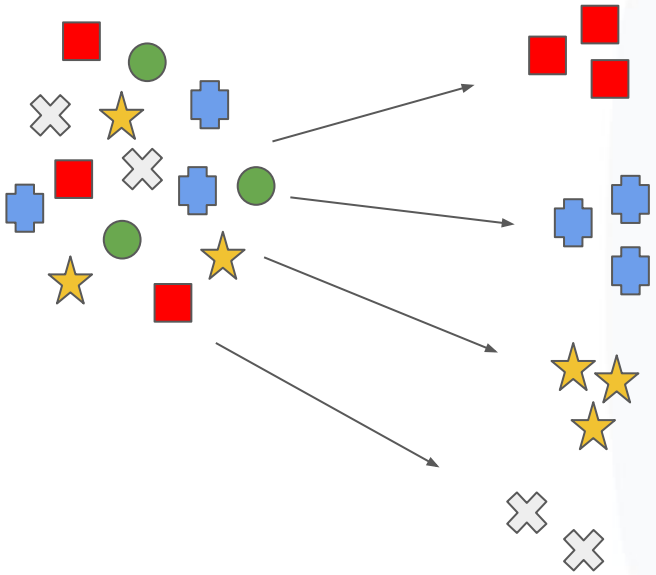
*Babelomics 5.0 based

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.

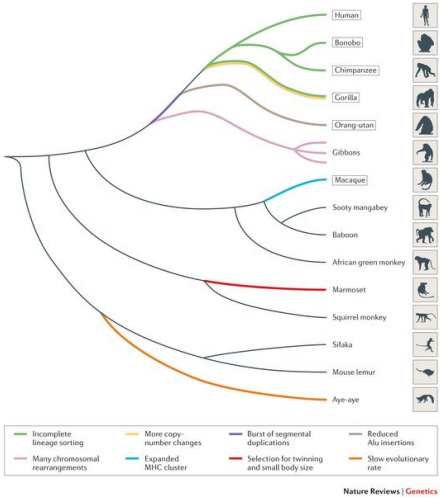
If plotted geometrically, the objects within the clusters will be close together, while the distance between clusters will be farther apart.



Applications of clustering analysis

Biology

Taxonomy of species



Social network analysis

Recognize communities within large groups of people.

Business and marketing

Product positioning, new product development, etc.

Beer and Nappies: Walmart, by using data mining discovered that by placing beers and nappies together increase the sale of both products.

Clustering analysis in molecular biology

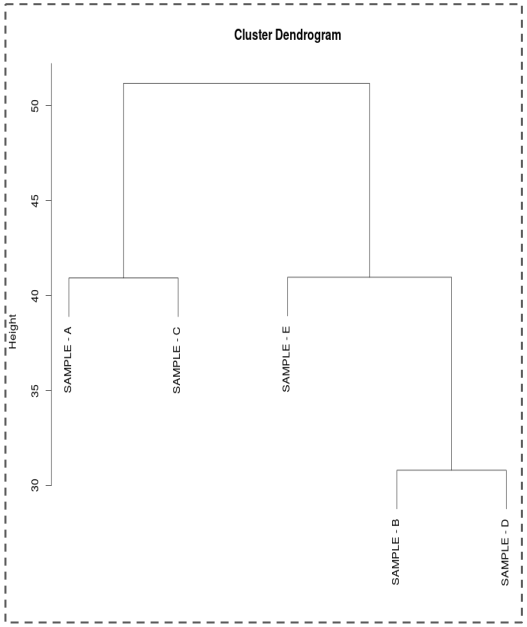
- In our case, we cluster **genes** and/or **samples.**

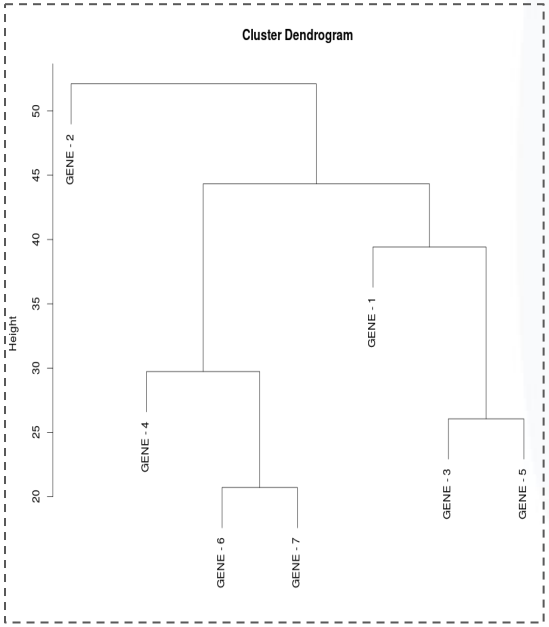
	sampleA	sampleB	sampleC	sampleD	sampleE
gene1	47	20	24	36	12
gene2	35	47	33	47	42
gene3	39	19	21	18	46
gene4	38	12	44	16	22
gene5	19	14	16	20	31
gene6	19	26	36	18	12
gene7	24	38	46	14	24

Select type of clustering: samples and/or genes

☐ Clustering of samples ☐ Clustering of genes

Options in Babelomics 5.0





Clustering analysis in molecular biology

Clustering is **unsupervised** classification.

- No predefined class.

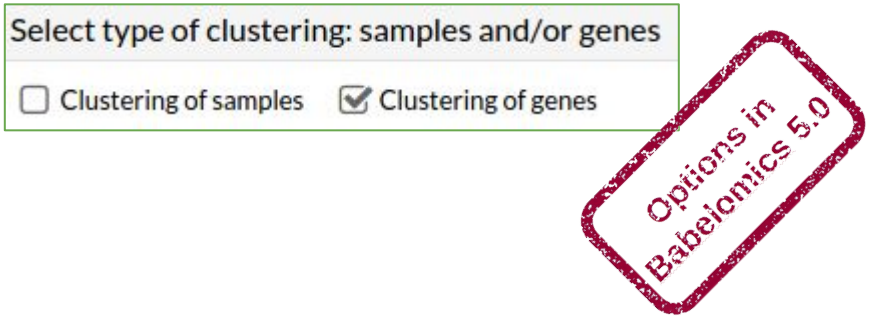
	sampleA	sampleB	sampleC	sampleD	sampleE
gene1	47	20	24	36	12
gene2	35	47	33	47	42
gene3	39	19	21	18	46
gene4	38	12	44	16	22
gene5	19	14	16	20	31
gene6	19	26	36	18	12
gene7	24	38	46	14	24

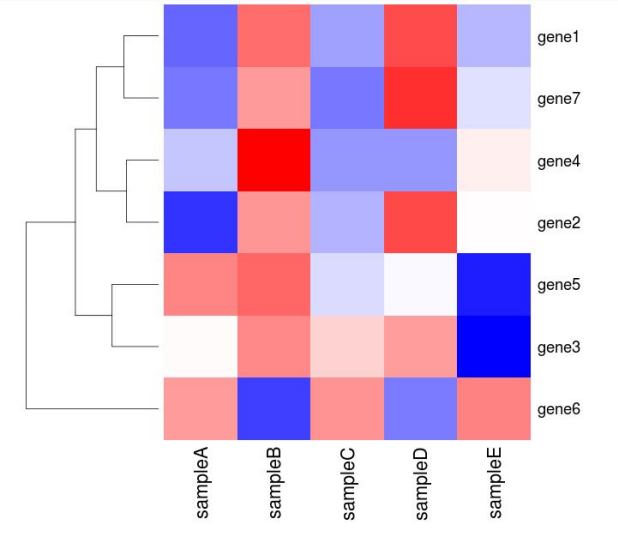
	sample_name	sample_type
1	"sampleA"	"Tumor"
2	"sampleB"	"Normal"
3	"sampleC"	"Tumor"
4	"sampleD"	"Normal"
5	"sampleE"	"Tumor"

Which questions can be answered with clustering?

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

- 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.

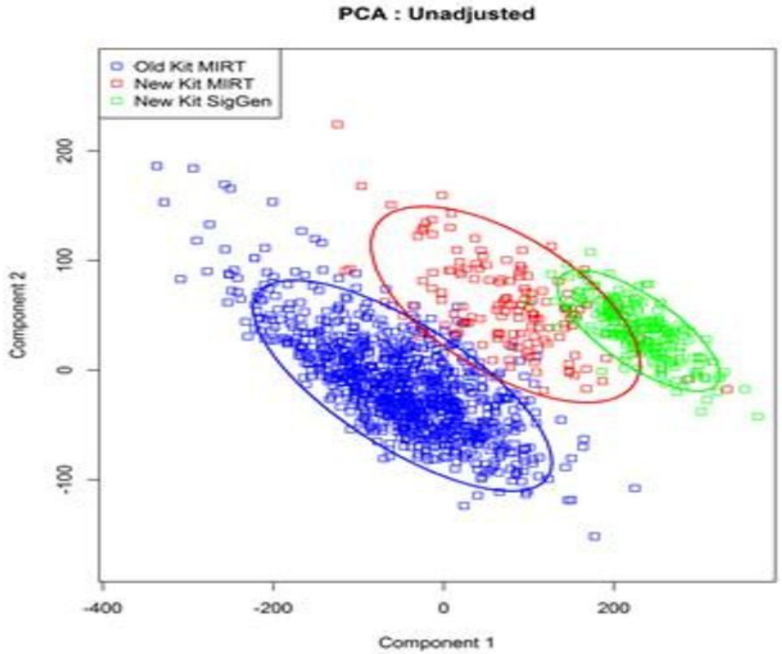


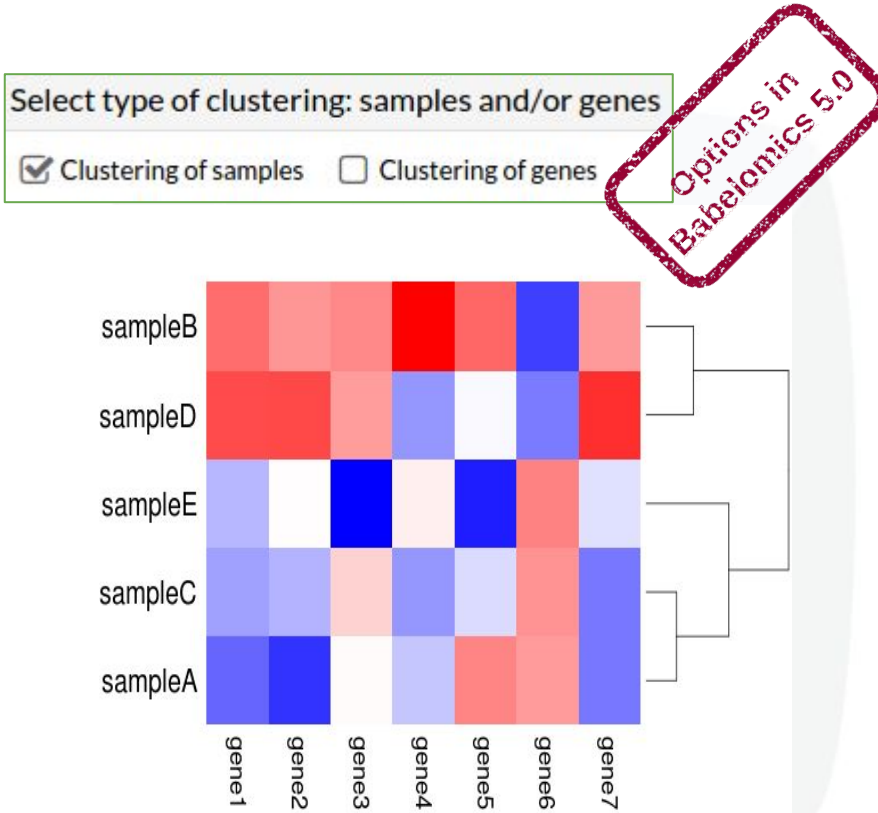


Which questions can we answer with clustering?

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

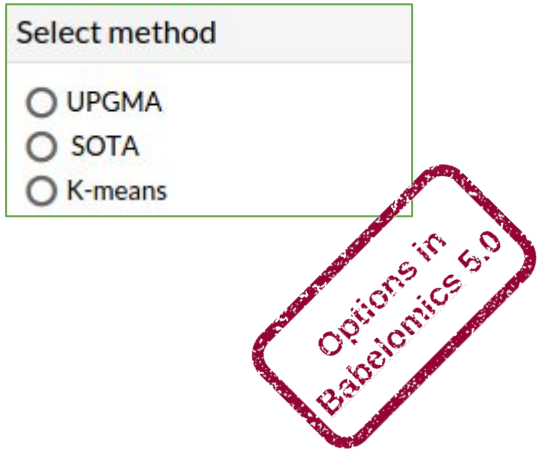
- The unit of analysis is the sample.
- Discover new subgroups in a set of patients of the same disease.
- Descriptive analysis.
- Perform quality control checking
  - Outlier detection
  - Batch effect assessment





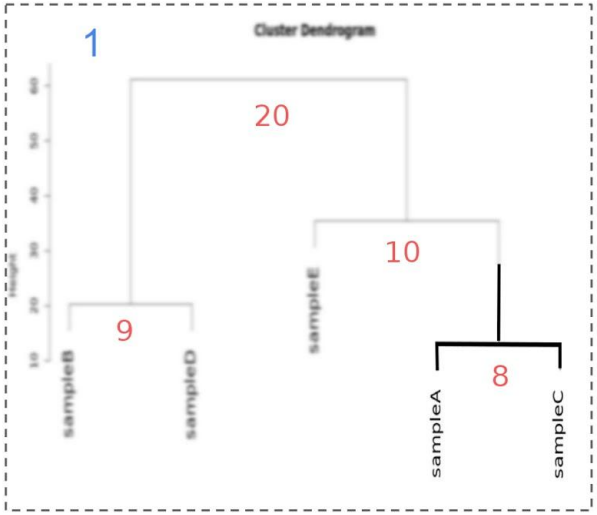
Clustering Methods in Babelomics 5.0

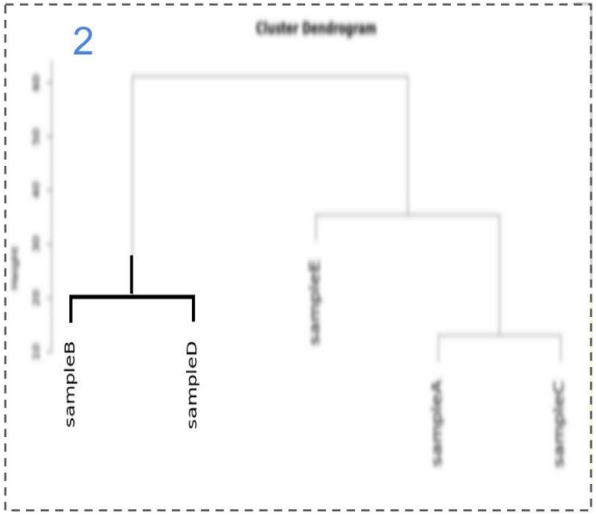
- Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
- Self-Organizing Tree Algorithm (SOTA)
- K-Means

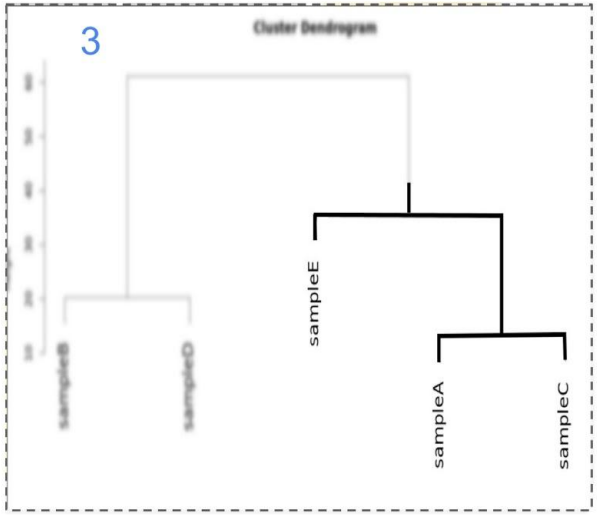


UPGMA

- 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.

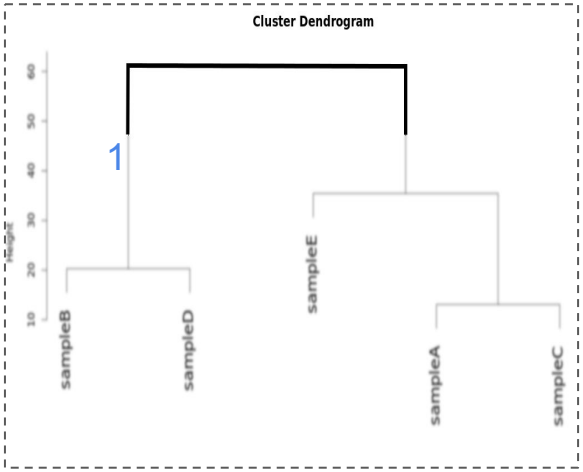


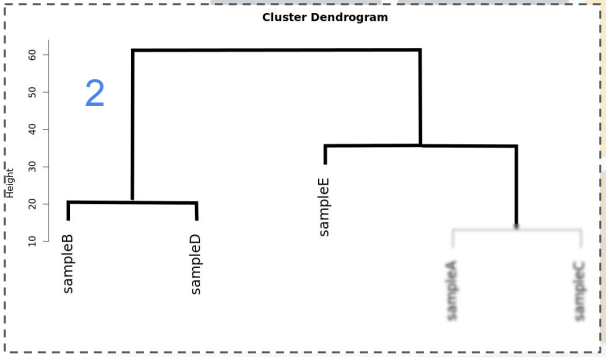




SOTA

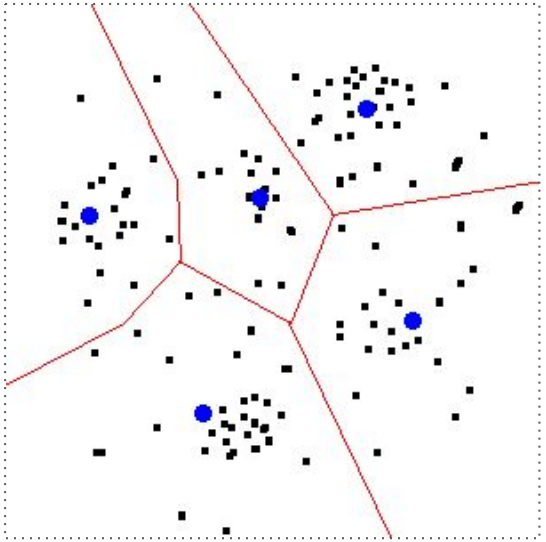
- SOTA starts the classification with a binary topology composed of a root node with two leaves.
- A divisive(top down) method.
- The self-organizing process splits the data (e.g. samples) into two clusters.
- Provides a tree.

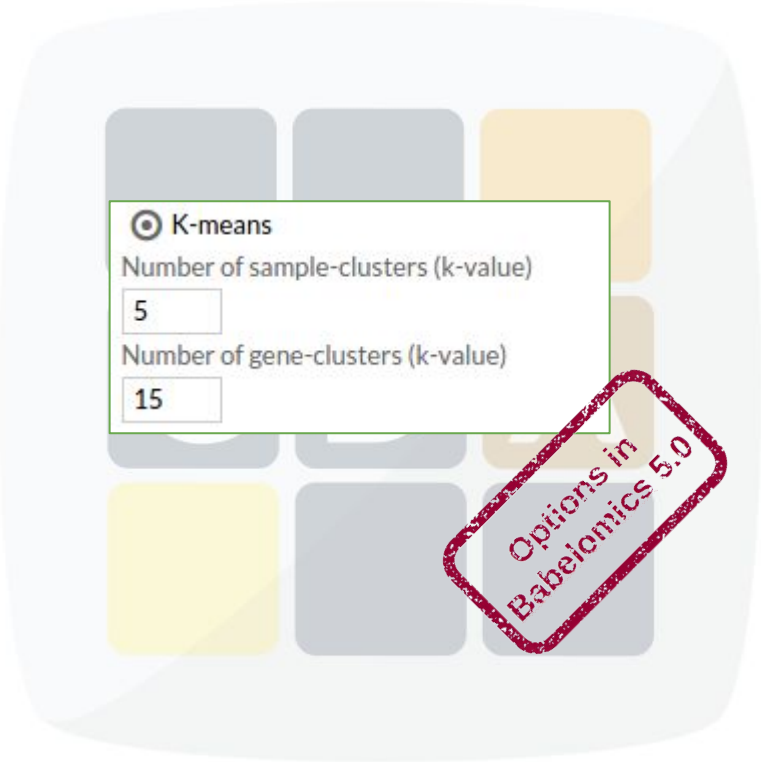




K-Means

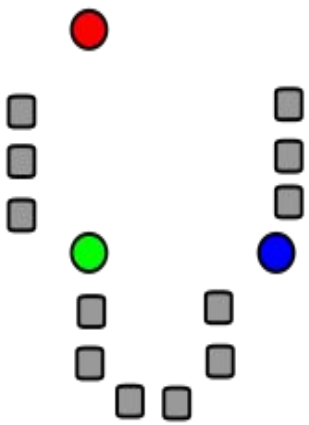
- K-means aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.
- Do not provide a tree.
- Usually need the number of cluster to be set.
- Its result is very sensitive to the initialization step: choosing initial cluster centers.





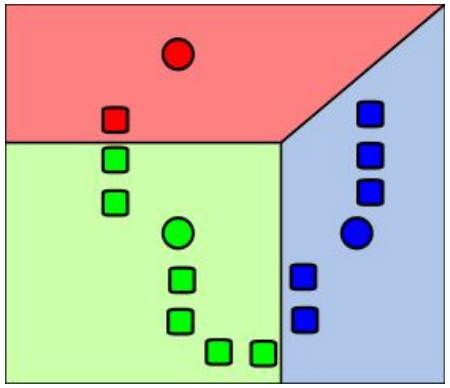
K-Means

1



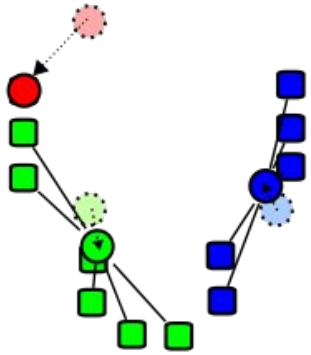
Red,Blue,Green circles are initial cluster centers.

2



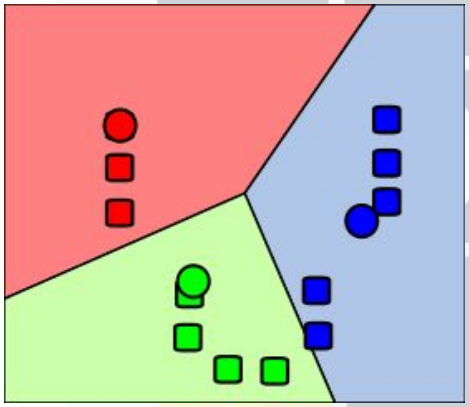
We find closest observations to initial centers.

3



We calculate center of new clusters.

4



Repeat step 2 and 3 until no changes occur.

Distance Parameters

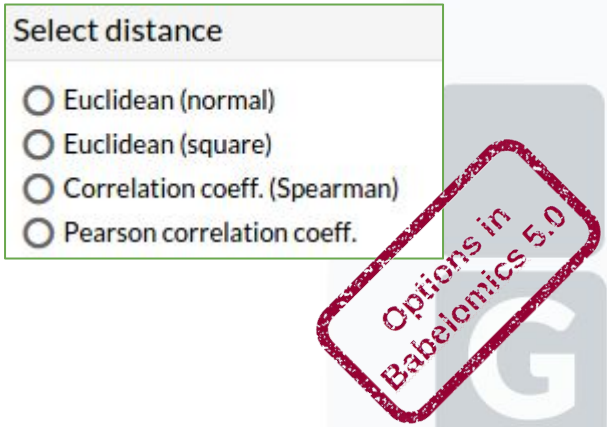
Different distances account for **different properties**.

1.Euclidean

- Normal
- Squared

2.Correlation coefficient

- Spearman
- Pearson



Distance Parameters

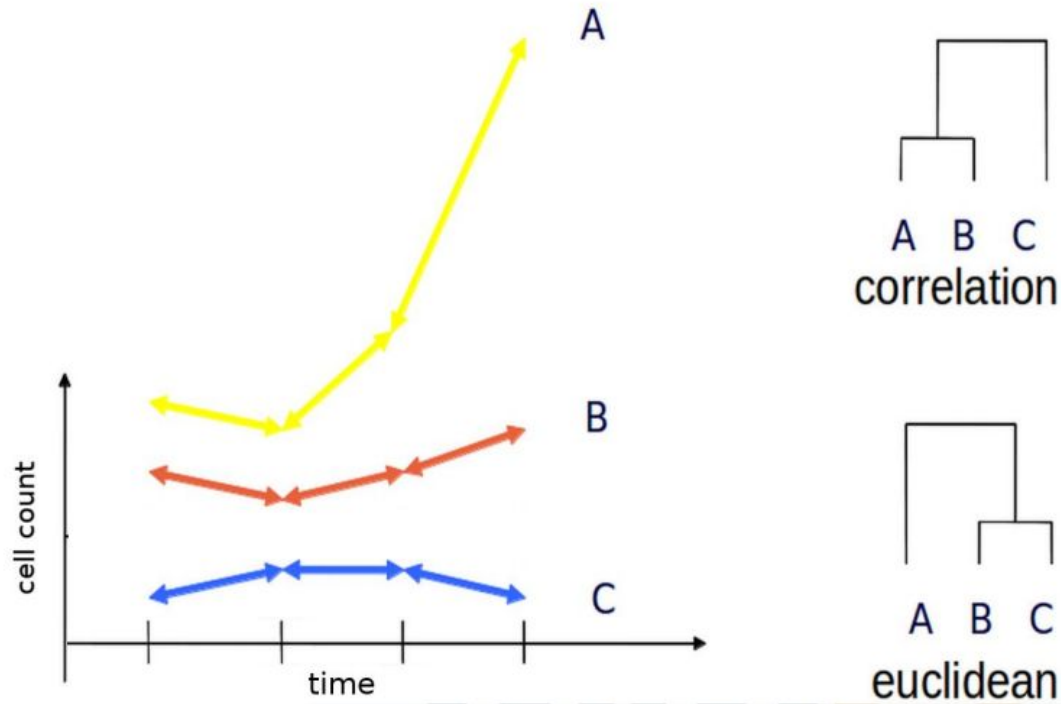
FAQs:

1. Which one is the best?
2. Which one I have to select?

Different definitions of **being close.**

Correlation: tendencies

Euclidean: global similarity

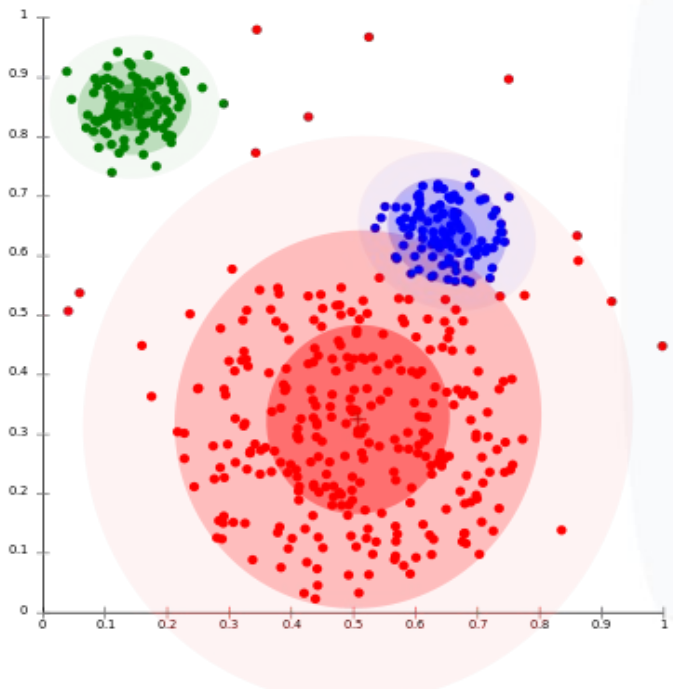


A good clustering

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

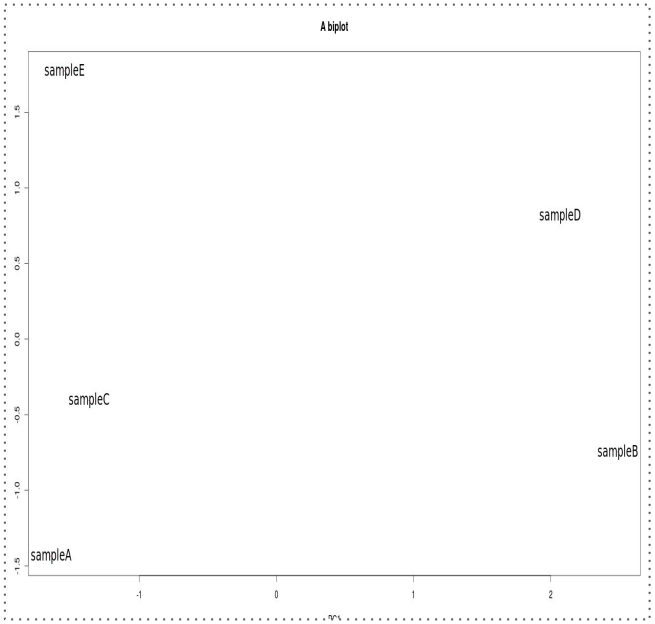
- High intra-class similarity (**Green**, **Blue** >> **Red**)
- Low inter-class similarity (**Green** vs. **Blue**, **Green** vs. **Red** >> **Blue** vs **Red**)

! Depends on the quality of data.

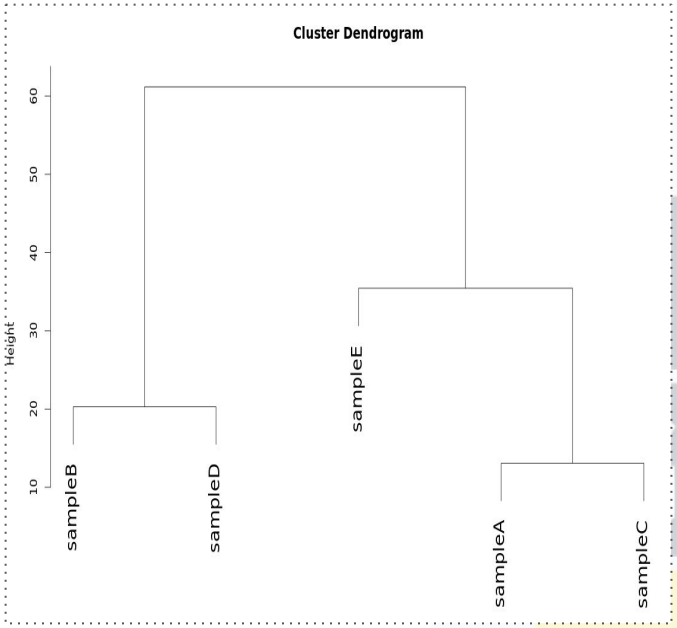


Visualisation of Clustering Results

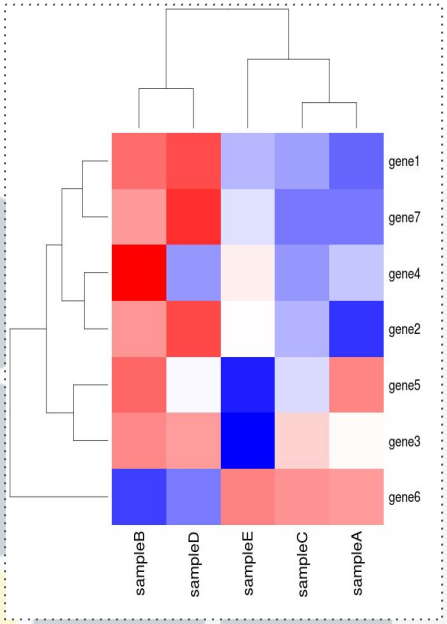
Biplot



Tree Diagram (Dendrogram)



Heatmap



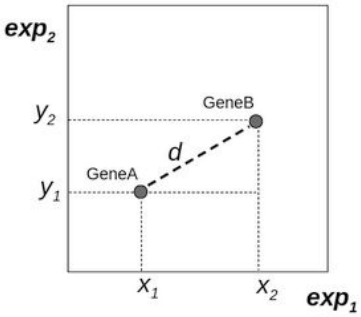
Examples: Distance Parameters

1. Euclidean

	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 Matrix	GeneA	GeneB
GeneA	1	12.85
GeneB	12.85	1

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



Examples: Distance Parameters

2. Correlation coefficient

- Pearson

	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

#Distance Matrix	GeneA	GeneB
GeneA	1 - 1 = 0	1 - 0.98 = 0.02
GeneB	1 - 0.98 = 0.02	1 - 1 = 0

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

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$ (simplified formula)

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

Examples: Distance Parameters

2. Correlation coefficient

- Spearman

	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	1 - 0.8 = 0.2
GeneB	1 - 0.8 = 0.2	1 - 1 = 0

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

n = number of samples

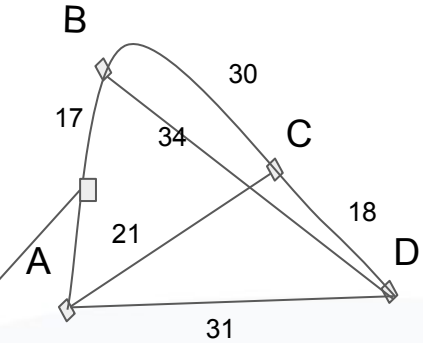
Cor coef. = 1 - (6*(1+1+0+0))/(4(16-1))

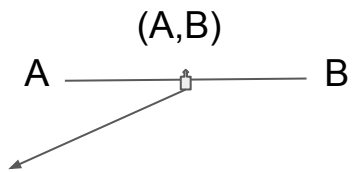
= 1 - (12/60)

= 0.8

Examples: UPGMA

#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.

Examples: UPGMA

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

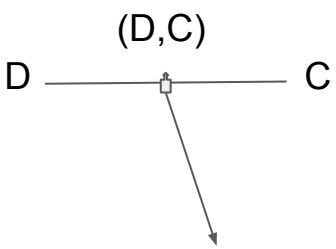
=(31+34)/2=32.5

#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	25.5	32.5
GeneC	25.5	0	18
GeneD	32.5	18	0

Examples: UPGMA

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

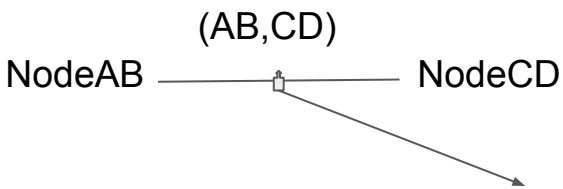
#Distance Matrix	NodeAB	GeneC	GeneD
NodeAB	0	22.5	32.5
GeneC	22.5	0	**18**
GeneD	32.5	**18**	0

#Distance Matrix	NodeAB	NodeCD
NodeAB	0	27.5
NodeCD	27.5	0

Examples: UPGMA

Third brach length estimation:

(Our root node)



#Distance Matrix	NodeAB	NodeCD
NodeAB	0	27.5
NodeCD	27.5	0

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

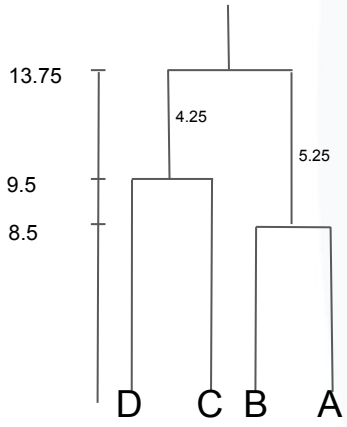
To create the dendrogram we deduce the other remaining branch lengths:

NodeRoot-NodeAB = 13.75 - 8.5

= 5.25

NodeRoot-NodeCD = 13.75 - 9.5

= 4.25



Supervised Classification Analysis
Babelomics 5.0

Cankut ÇUBUK
March 9th, 2017

GDA
International Course on
Genomic Data Analysis

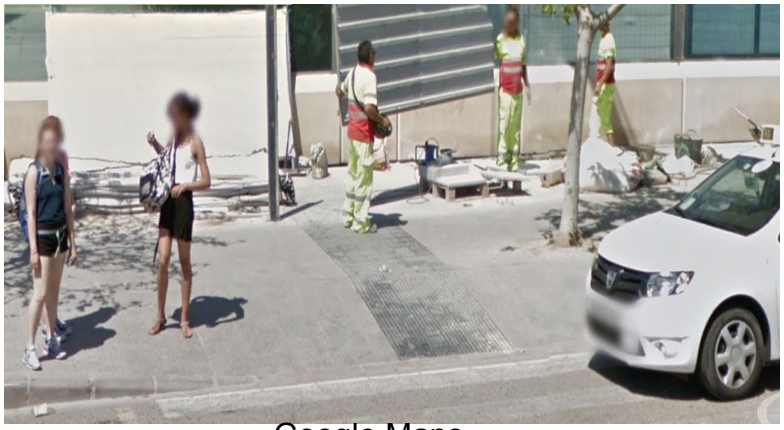
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Outline

- Brief history
- Introduction
- Supervised classification methods in Babelomics 5.0
- Error estimation of classification
- Feature selection

Brief History

- Predictors/classifiers are a subset of methods of the Artificial Intelligence area.
- 1956, John McCarthy coined the term “Artificial Intelligence” and defined it as “the science and engineering of making intelligent machines.”
- 60-70's, the mathematical background of some of these algorithms/methods were developed.
- 70-80's, with the apparition of computers first predictors were developed for many different areas:
    - handwriting recognition
    - weather prediction
    - face recognition
    - speech recognition



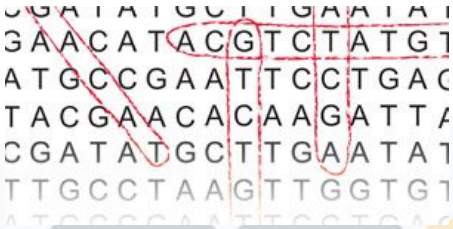
Google Maps

Prisma App

Facebook

Brief History

- 90's, predictors begun to be used in bioinformatics and computational biology:
  - genome annotation: sequence based TFBS, gene, intron, exons prediction/finding
  - protein structure prediction

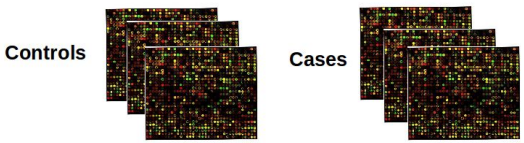


- In late 90's, DNA microarray technology was developed

- In early 2000, two questions arose:
  - Could biological samples be classified according to their gene expression values?
  - Could we use computers to classify these samples?

Brief History

- 2002, appears the first paper applying predictors method to DNA microarray data, van't Veer et al..

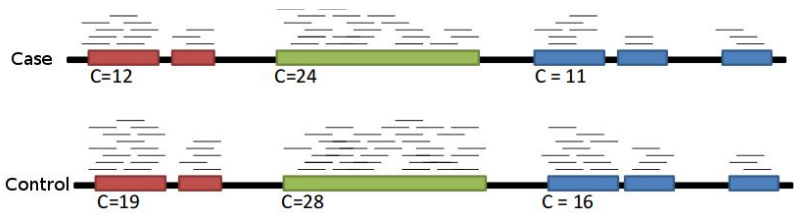


Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

NATURE | VOL 415 | 31 JANUARY 2002 | www.nature.com

- Since that moment hundreds of papers, applications and new methods have been developed which are also used for NGS data (e.g. RNAseq).



VOLUME 27 · NUMBER 8 · MARCH 10 2009

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes

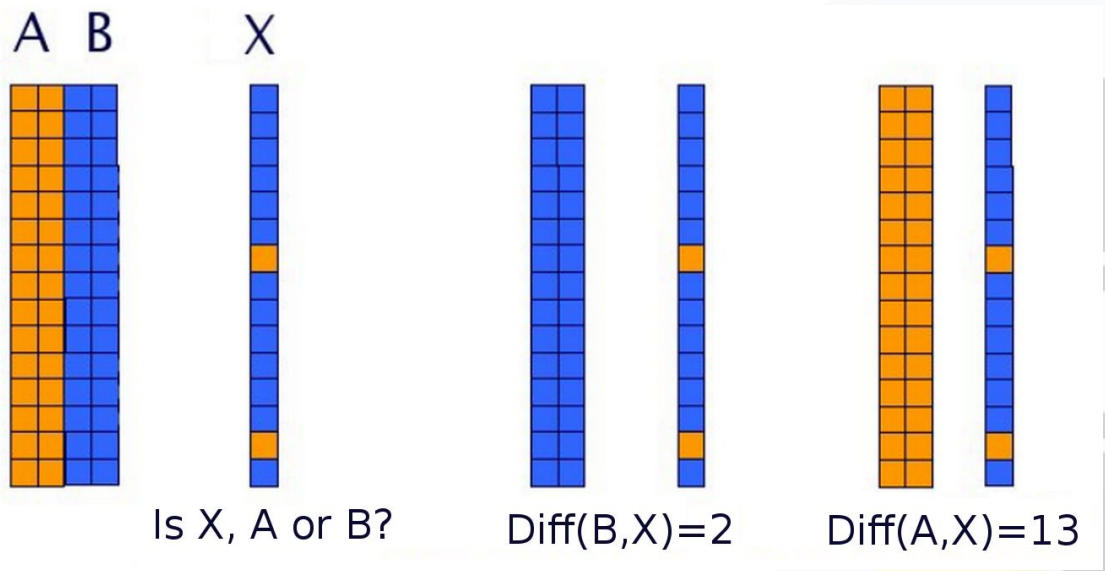
Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J.S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard

A B S T R A C T

Purpose
To improve on current standards for breast cancer prognosis and prediction of chemotherapy benefit by developing a risk model that incorporates the gene expression–based "intrinsic" subtypes luminal A, luminal B, HER2-enriched, and basal-like.

Creating Predictor (Training)

- Predictors need to be trained.
- Set of algorithms/methods that are going to allow the computer to **learn** a specific **labelled** problem and then be able to predict or classify new unlabelled samples is called supervised classification.



Training

Known Class Data (*Babelomics format*)

Class Label

#VARIABLE *tumor* CATEGORICAL{c1,c2,c3} VALUES{c1,c1,c2,c2,c3,c3}

GenesIDs / ProbeIDs

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116
1487_at	7.646126117	7.544048497	8.754540699	8.476873549	9.084035203	9.028724488
1494_f_at	7.498031252	7.679595836	7.662561072	7.201093115	7.426192546	7.669669586
1598_g_at	10.31770877	10.92530764	10.50092321	9.630201704	10.23473332	10.49766918
160020_at	8.529411037	8.738065073	8.617216353	8.445386532	8.425365655	8.76023381
1729_at	9.607320487	8.171988017	8.73040537	8.978602862	9.156752025	8.033237589
1773_at	6.216319215	6.441555855	6.165785507	6.325464779	6.121753223	6.229420354
177_at	6.535525364	6.453887146	6.519400663	6.333366799	6.385077422	6.407541976

arrays

Unknown Class Data

which is the class label?

1007_s_at	11.28578155
1053_at	7.787503325
117_at	7.487539205
121_at	9.489979282
1255_g_at	5.000099854
1294_at	8.358097049
1316_at	7.187245349
1320_at	5.645994428
1405_i_at	7.138444163
1431_at	4.697298725
1438_at	7.430761532
1487_at	8.646126117
1494_f_at	7.498031252
1598_g_at	10.31770877
160020_at	8.529411037
1729_at	9.107320487
1773_at	6.216319215
177_at	6.535525364

Select train data

Options in Babelomics 5.0

Select test data (Optional)

Algorithms in Babelomics 5.0

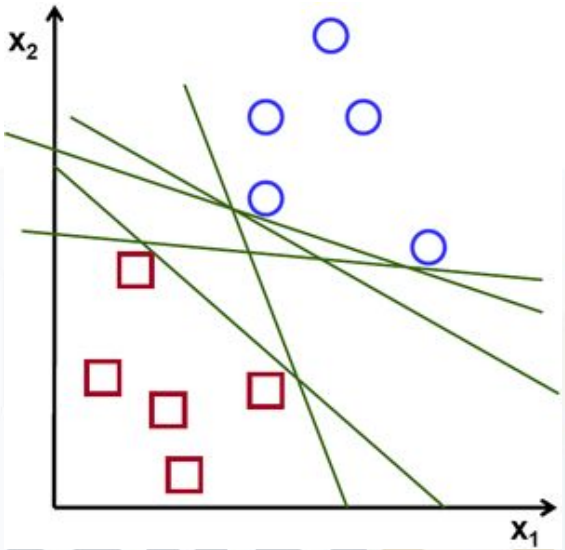
- Support Vector Machine (SVM)
- k-Nearest Neighbors (KNN)
- Random Forest

SVM

- A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.

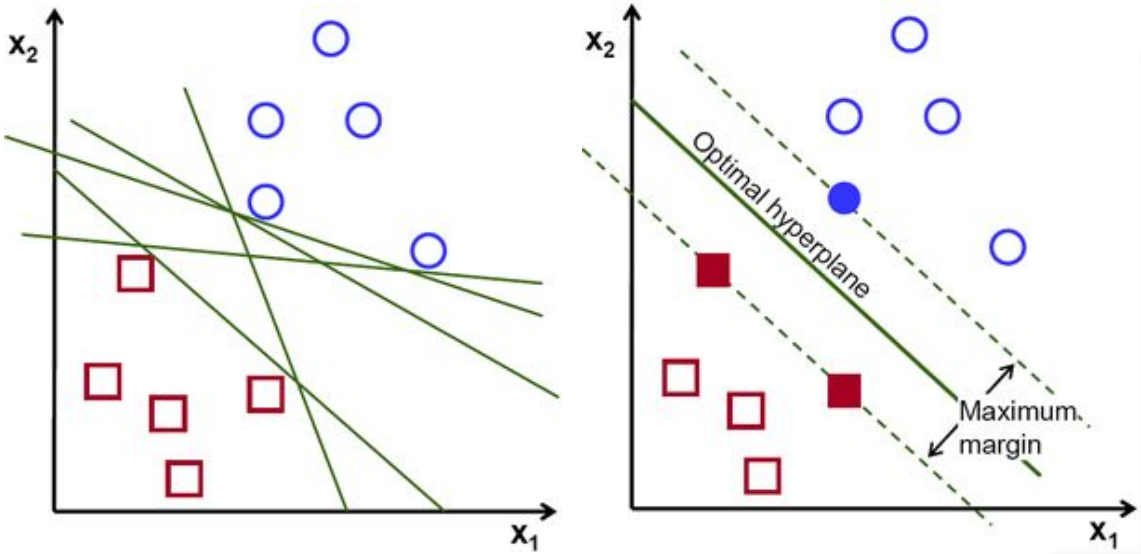
- In the figure, you can see that there are multiple lines that offer a solution to the problem.

- Is any of them better than the others?



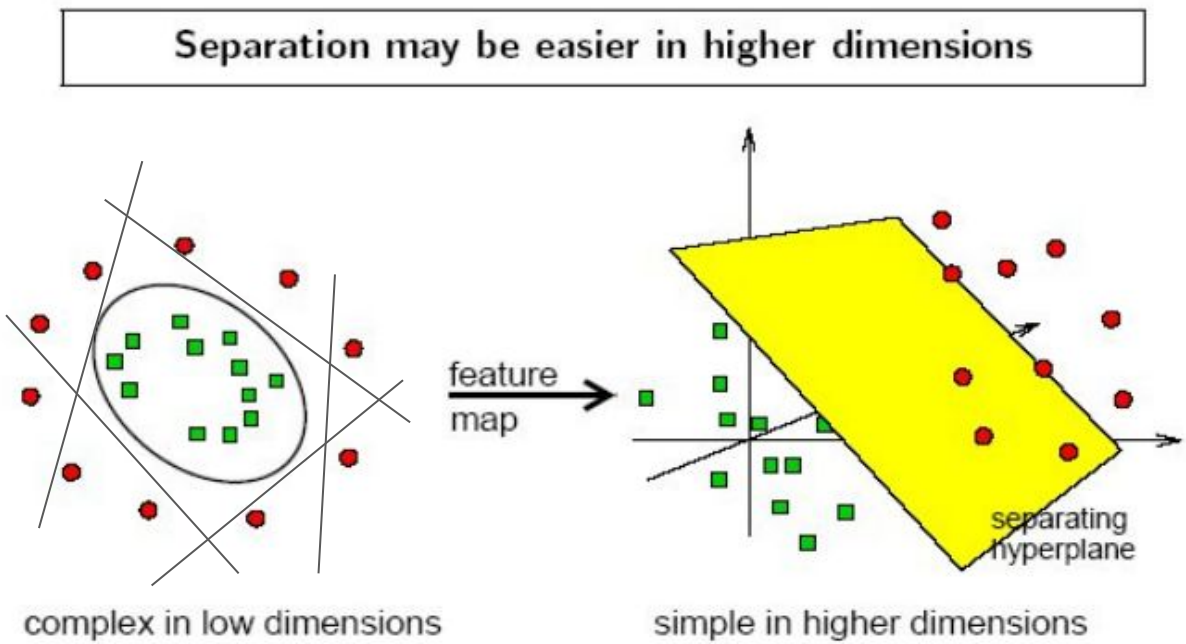
SVM

A special property of SVMs is that maximizes the margin between the decision hyperplane and the training samples.



SVM

- But many times the data does not have a linear solution.
- Then we can use a **kernel trick** and map the data into a higher dimension space.

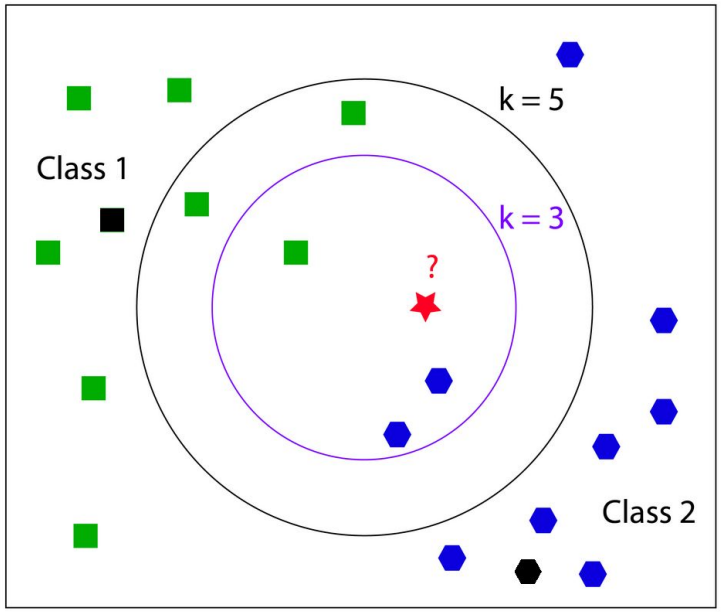


KNN

- k-Nearest Neighbor (KNN) is a distance based prediction method.

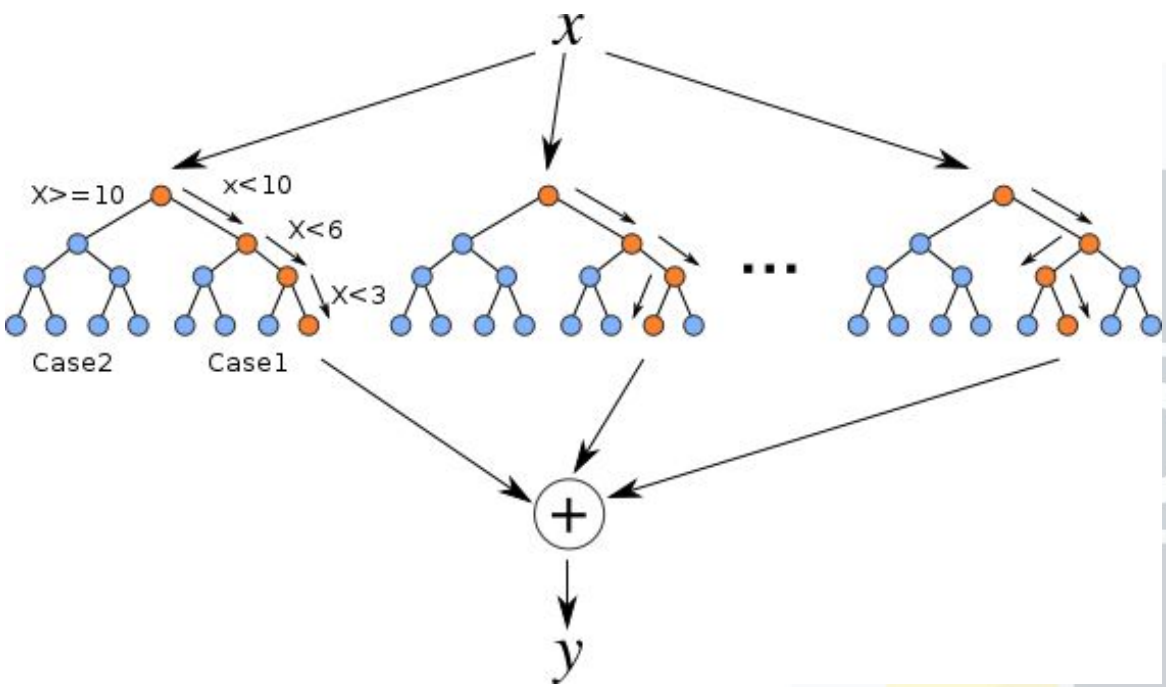
- In order to make KNN more robust we are going to look for the K nearest neighbours instead of only the nearest.

- Predictor tool test automatically with **K=1..20**



Random Forest

- Random Forest builds many trees using a subset of the available input variables and their values.
- The forest chooses the classification having the most votes over all the trees in the forest.



Feature selection

Feature selection is the technique of selecting a subset of relevant features for building robust learning models.

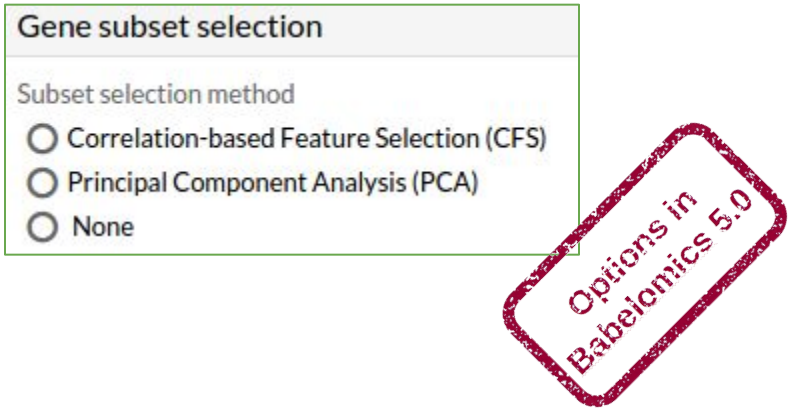
- Finds the genes which discriminate classes.
- Removes the genes which change randomly within classes.
- Genes that do not change will not bear any information and not useful for classifying.
- Increases the accuracy.
- Decreases analysis time.

#VARIABLE ***tumor*** **CATEGORICAL{c1,c2,c3} VALUES{c1,c1,c2,c2,c3,c3}**

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116
1487_at	7.646126117	7.544048497	8.754540699	8.476873549	9.084035203	9.028724488
1494_f_at	7.498031252	7.679595836	7.662561072	7.201093115	7.426192546	7.669669586
1598_g_at	10.31770877	10.92530764	10.50092321	9.630201704	10.23473332	10.49766918
160020_at	8.529411037	8.738065073	8.617216353	8.445386532	8.425365655	8.76023381
1729_at	9.607320487	8.171988017	8.73040537	8.978602862	9.156752025	8.033237589
1773_at	6.216319215	6.441555855	6.165785507	6.325464779	6.121753223	6.229420354
177_at	6.535525364	6.453887146	6.519400663	6.333366799	6.385077422	6.407541976

Feature selection

- Correlation-based Feature Selection (CFS)
- Principal Components Analysis (PCA)



Error Estimation of Prediction

It is not a simple task, we have to estimate the error that the predictor will have in future gene expression data.

This estimation can only be done during the training stage.

- Leaving-one-out cross-validation
- k-fold cross-validation

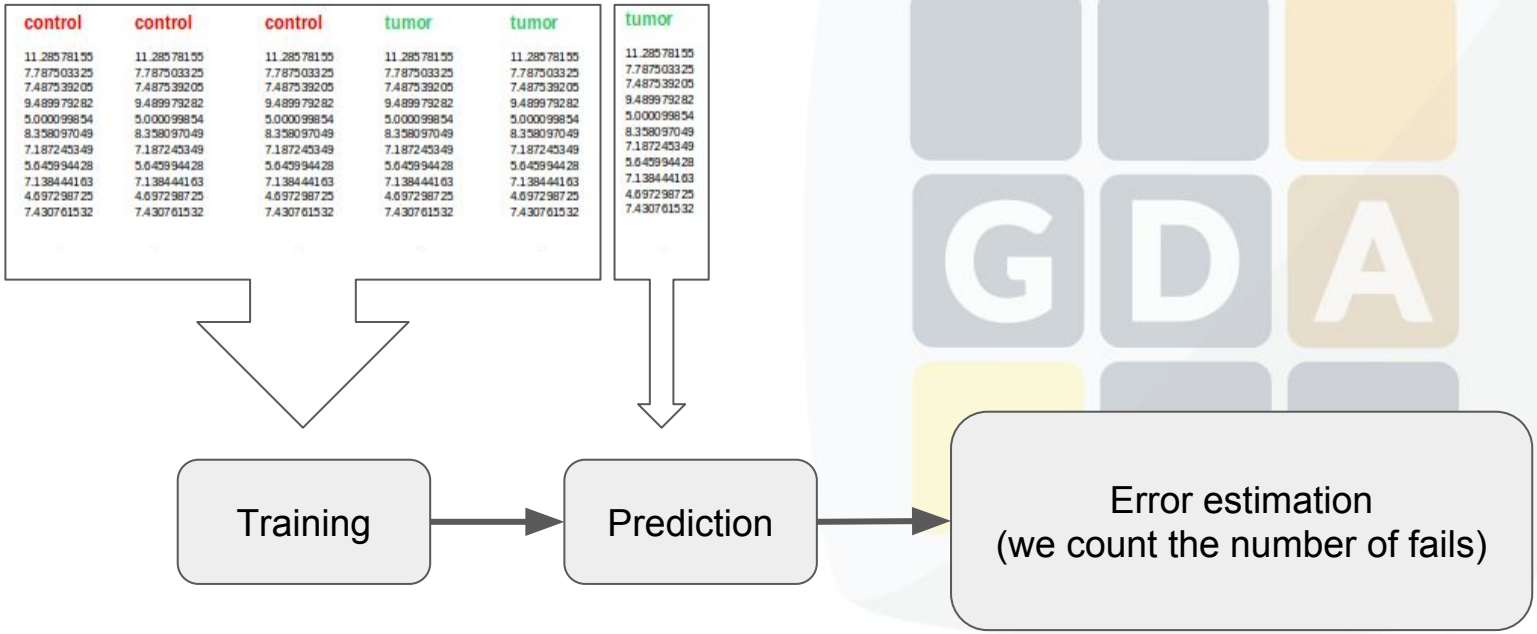
Leaving-one-out cross-validation

We use

- one sample to use as a test set.
- the rest as a training set.

k = number of arrays i.e.: k=6

Repeat k times changing the array to be used in prediction



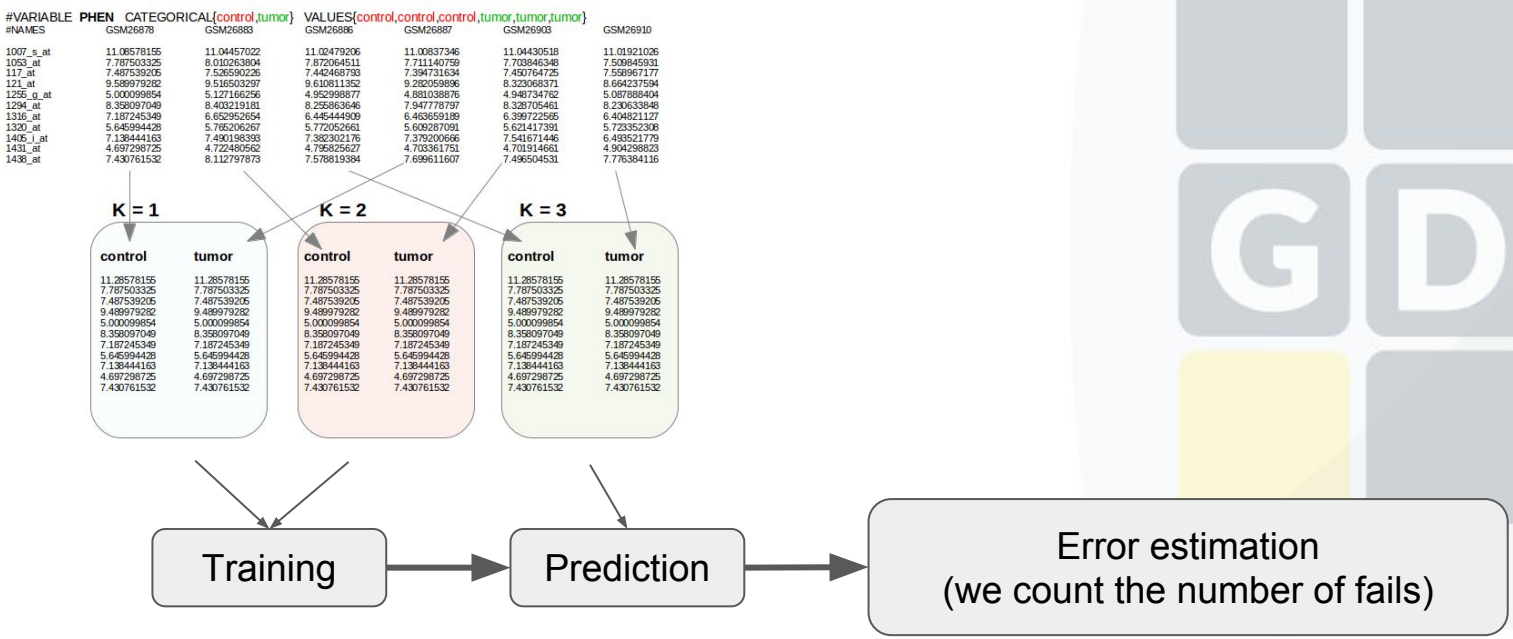
k-fold cross-validation

With this method we are going to split the data in k partitions.

We use

- (k-1) partitions as a training set.
- 1 partition as a test set.

In the example below, we split arrays into k=3 partitions of equal size.



Error metrics

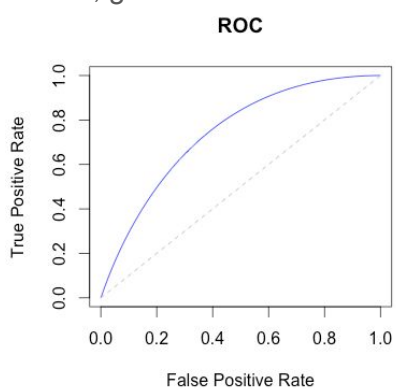
Confusion matrix

TP **(True Positive)**	**FP** **(False Positive)**	*P** *(Total Predicted Positive)*
FN **(False Negative)**	**TN** **(True Negative)**	*N** *(Total Predicted Negative)*
P *(Total Positive)*	*N* *(Total Negative)*	D (Total)

- **Accuracy (ACC)**
  bad:0, good=1

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N}$$

- **Area Under ROC (AUC)**
  bad:0, good=1



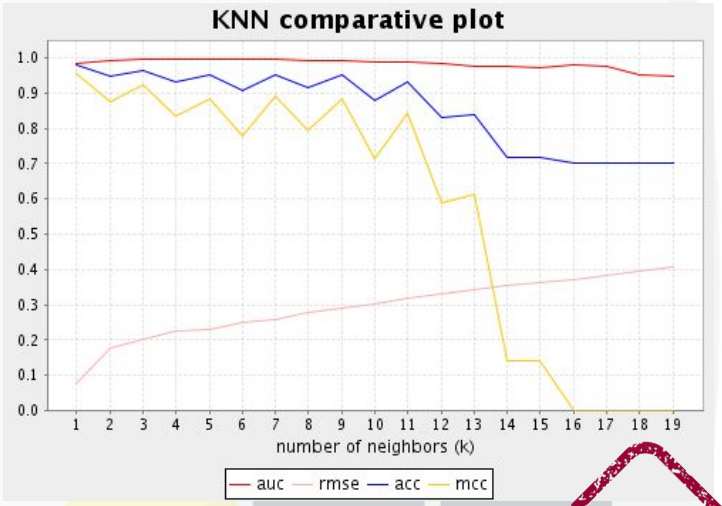
- **Matthews correlation coefficient (MCC)**
  bad:0, good=1

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

- **Root Mean Square Error (RMSE)**
  bad:1, good=0

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}.$$

y_t: real
$\hat{y}_t$: predicted



Exercises on Babelomics

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