**C. Classification**

Principio del formulario

Final del formulario

**C.1. Unsupervised classification**

How can we detect groups of patients with similar expression profile? What microRNAs or genes have a common intensity pattern for an experimental group? Could we explore our data before continuing the analysis?

**Activity 1.** [**Online example**](http://bioinfo.cipf.es/WODA/doku.php/unsupervised.class/exercises/ej1)

1. Go to the Babelomics page and select the Clustering option from the Expression menu.
2. Press the online example and you will see how the parameters and form fields are now filled. As you can notice, this example is prepared to perform a clustering analysis on genes (rows) and conditions (columns) using the K-means algorithm with 5 sample-clusters and 15 gene-clusters. Here, the selected distance is Euclidean (square).
3. Press Launch job, and wait for your job to be finished.
4. When the process finishes, a new blue job is shown at the right side of the web page. Press it to check your results.

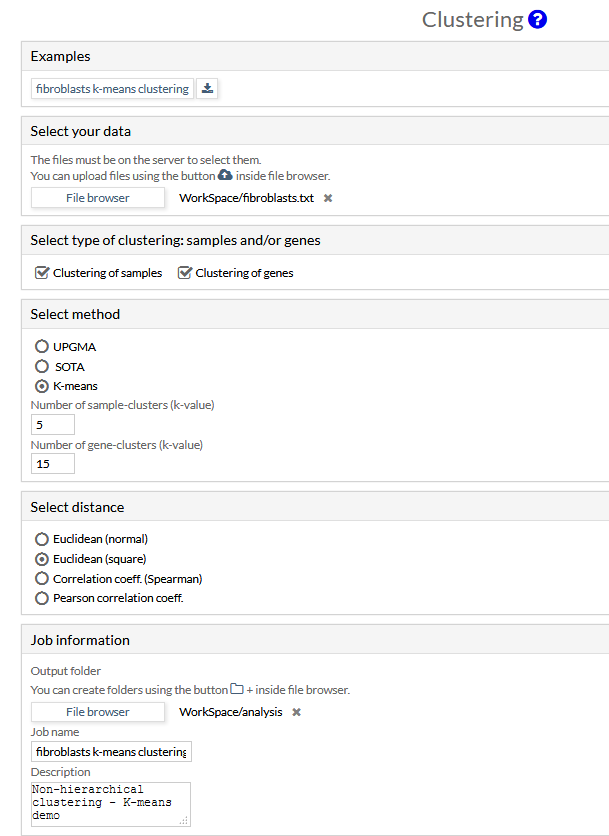
Principio del formulario

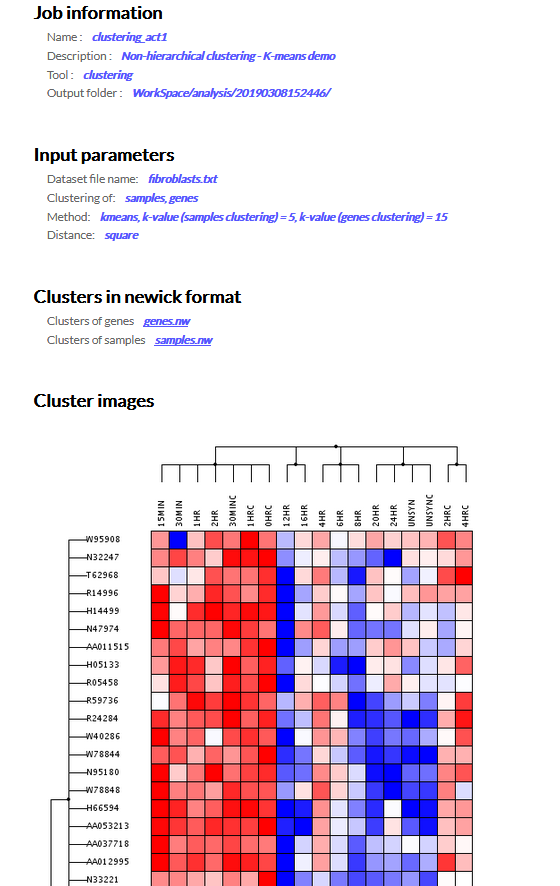
Final del formulario

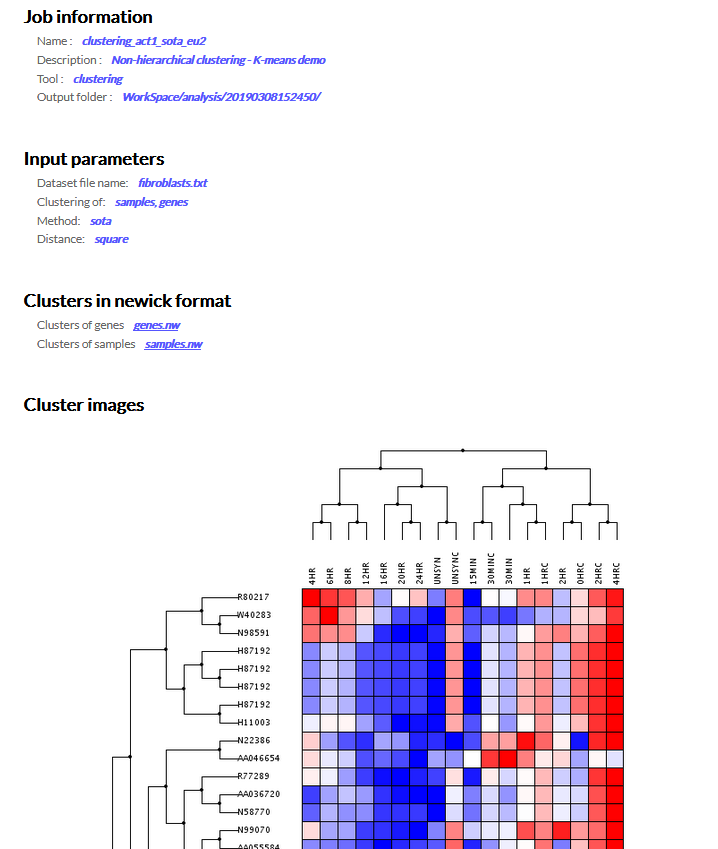
## Questions

These are some questions that you should be able to answer about the previous example:

1. Do you think that the clustering was able to differentiate any group of coexpressed genes?
2. How many sample clusters are there? and gene clusters?
3. Launch this online example using different clustering methods and compare the results. Which are the differences between the results of these results for different methods?
4. What about newick format?







## **ACTIVITY 2.**

## **Clustering analysis for expression data in arthritis**

The etiology of **rheumatoid arthritis** is not known with certainty. In order to generate information that clarifies this point, a study of expression microarrays has been proposed, which will allow characterizing this disease at the molecular level and finding some key mechanisms that will improve its prevention and treatment.

### Goal

Detect homogeneous groups of subjects according to their transcriptomic profile and evaluate the possible presence of anomalous patterns.

Principio del formulario

Final del formulario

### Data

We have normalized data from Affymetrix microarrays for three experimental groups:

* 5 patients with rheumatoid arthritis (RA1-RA5).
* 4 patients with osteoarthritis (OA1-OA4).
* 6 healthy people (H1-H6).

**Work plan**

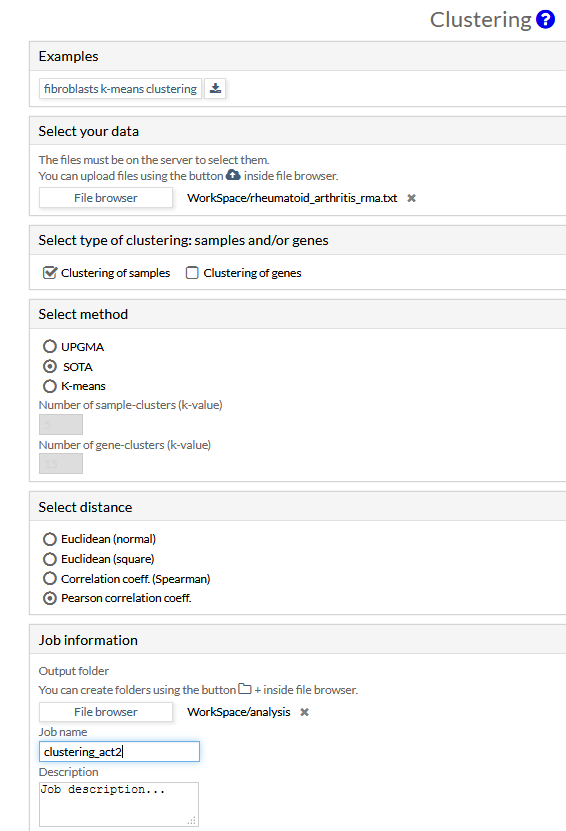
1. Open the data file of [**gene expression**](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/unsupervised.class/exercises/rheumatoid_arthritis_rma.txt) with a spreadsheet and inspect its contents. There will be as many columns as subjects and as many rows as genes.
2. Upload this txt file in Babelomics from the “Upload” menu. We will have to indicate the type of data that we upload: “Data matrix expression”. This link describes the different types of data that we can use in Babelomics: <https://github.com/babelomics/babelomics/wiki/Data-types>.
3. Next, we select the clustering by samples. We chose the “SOTA” clustering method and the distance “Pearson correlation coefficient”. We assign a name to the job and execute it.
4. Perform a clustering for genes (to begin with, those that are by default). We assign a name to the job and execute it.

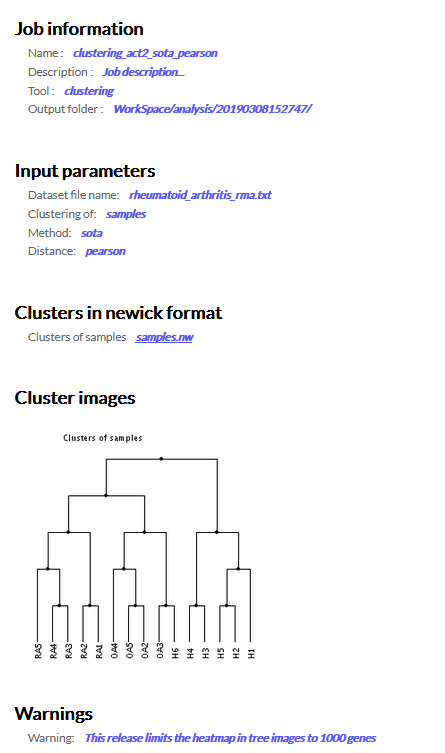
Principio del formulario

Final del formulario

**Questions**

1. Are there groups of samples with a similar transcriptomic profile? How many groups appear?
2. Is there any sample that has an anomalous behavior when comparing with other subjects? Any proposal?
3. Do you think that if we performed a differential expression analysis we would obtain a large number of differentially expressed genes?
4. Any incidence with clustering by genes?





## ACTIVITY 3. RNA-Seq data analysis: unsupervised classification or clustering

Principio del formulario

Final del formulario

### Goal

Detect homogenous groups of subjects according to their transcriptomic profile.

Principio del formulario

Final del formulario

### Data

We are studying a complex disease in which we know that a certain hormone has an important role. For them, we designed an experiment with RNA-Seq in mice with two groups: 6 wild type mice (WT) and 6 mice treated with T3 hormone.

These data were obtained after applying a primary analysis that included the evaluation of the quality of the sequences, mapping and quantification of expression at the gene level. We have expression levels (non-normalized counts) for the 12 mice described in 38,293 genes.

Principio del formulario

Final del formulario

### Work plan

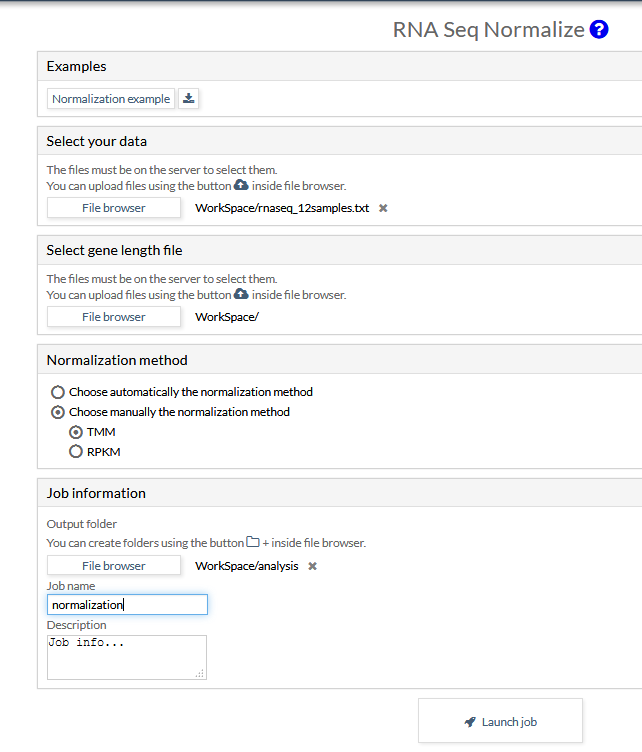
1. Open the data file of [**rnaseq\_12samples.txt**](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/unsupervised.class/exercises/rnaseq_12samples.txt) with a spreadsheet and inspect its contents. There will be as many columns as subjects and as many rows as genes.
2. Upload this txt file in Babelomics from the “Upload” menu. We will have to indicate the type of data that we upload: “Data matrix expression”. This link describes the different types of data that we can use in Babelomics: <https://github.com/babelomics/babelomics/wiki/Data-types>.
3. After loading the data, the first step will be normalization. From “Processing / Normalization NGS: RNA-Seq” we will select our file and choose a standardization method (we will start with TMM). Interesting clue: when the normalization finishes, check out the results and in the “Job information” section, look up the identifier of the “Output folder”. Then we will need it to indicate to Babelomics where are the normalized data.
4. Once the data is already normalized, we are ready to perform the clustering. From “Expression / Unsupervised analysis”, select the data (now it's time to select the previous “output folder” where the normalized data are ready).
5. Next, we select the clustering by samples. We chose a method of clustering and distance (to begin with, those that are by default). We assign a name to the job and run it.
6. Perform a clustering for genes (to begin with, those that are by default). We assign a name to the job and execute it.

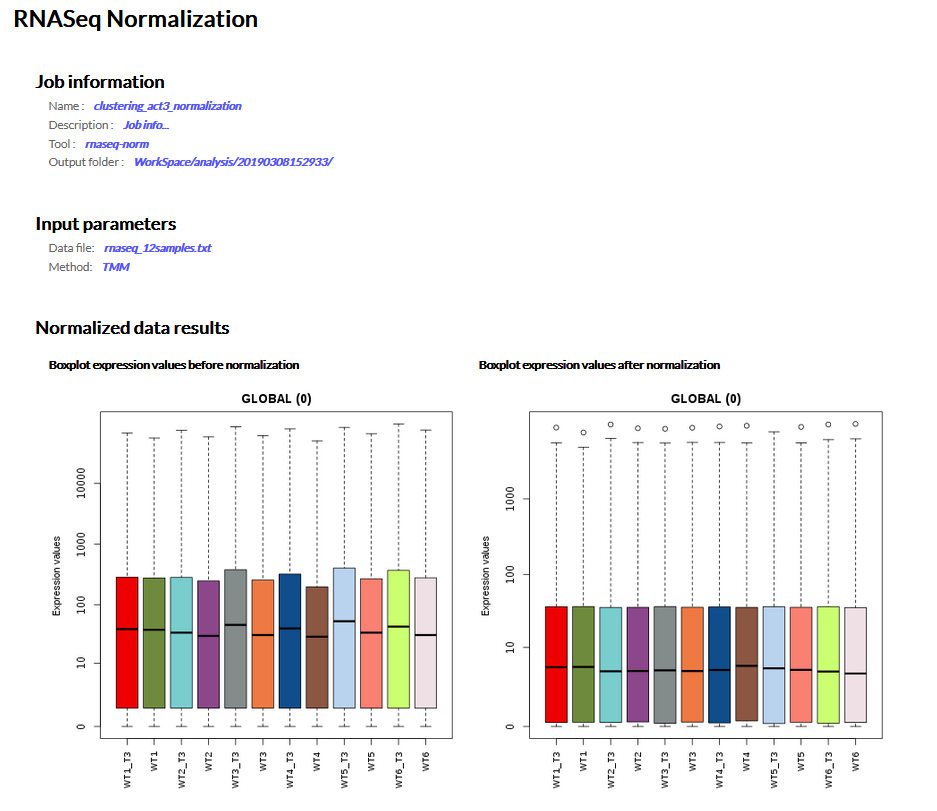
Principio del formulario

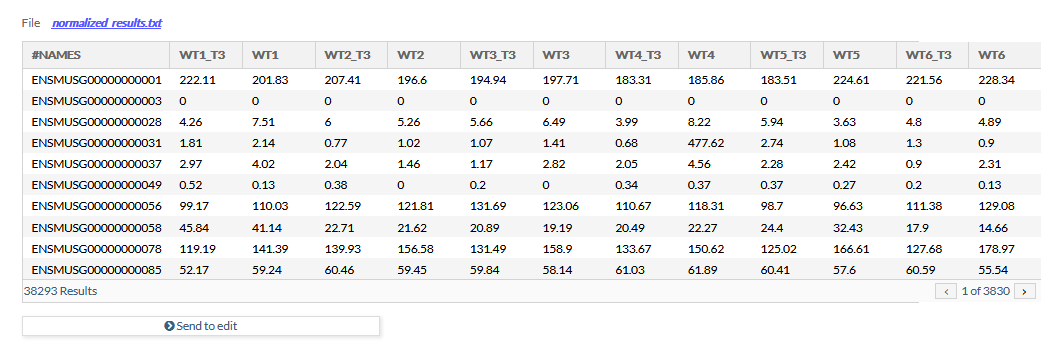
Final del formulario

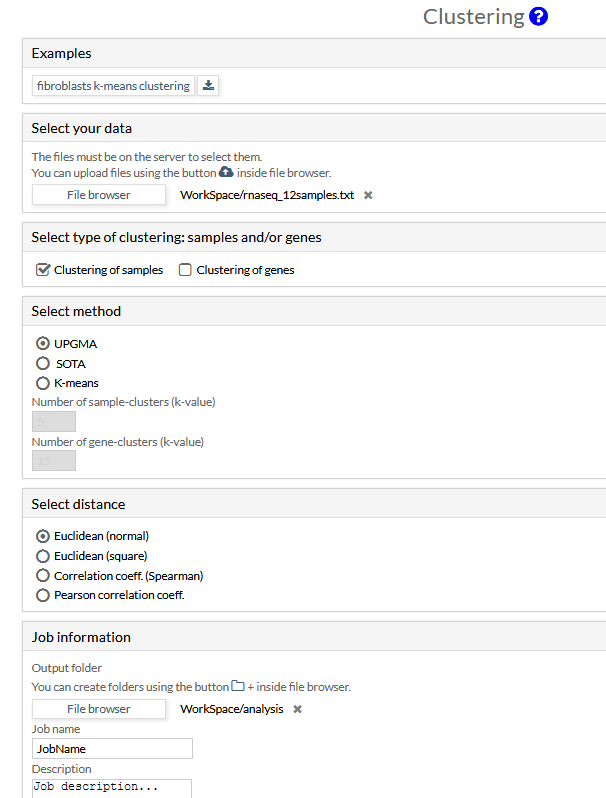
### Questions

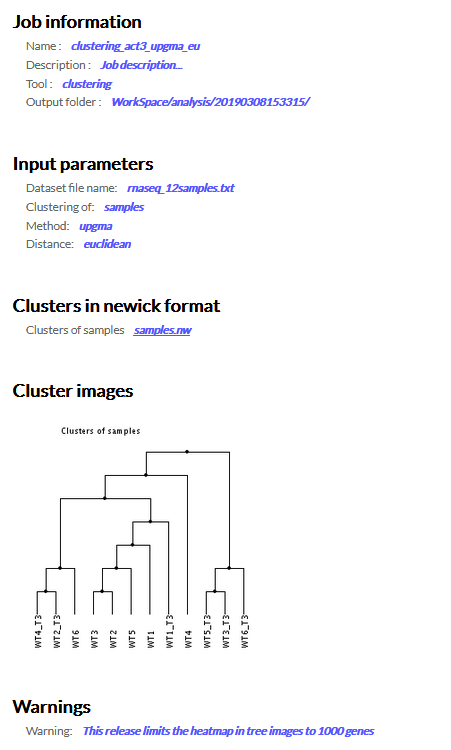
1. Are there groups of samples with a similar transcriptomic profile? How many groups appear?
2. Is there any sample that has an anomalous behavior when comparing with other subjects?
3. Do you think that if we performed a differential expression analysis we would obtain a large number of differentially expressed genes?
4. Any incidence with clustering by genes?

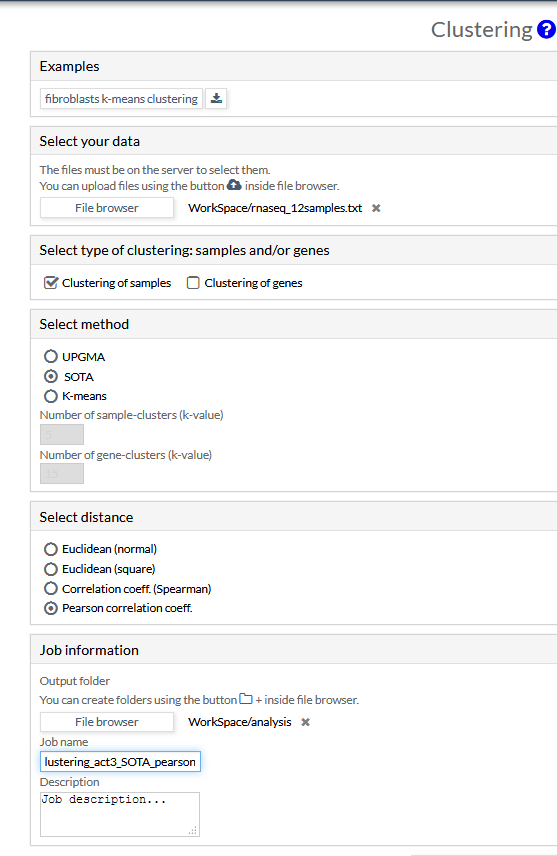


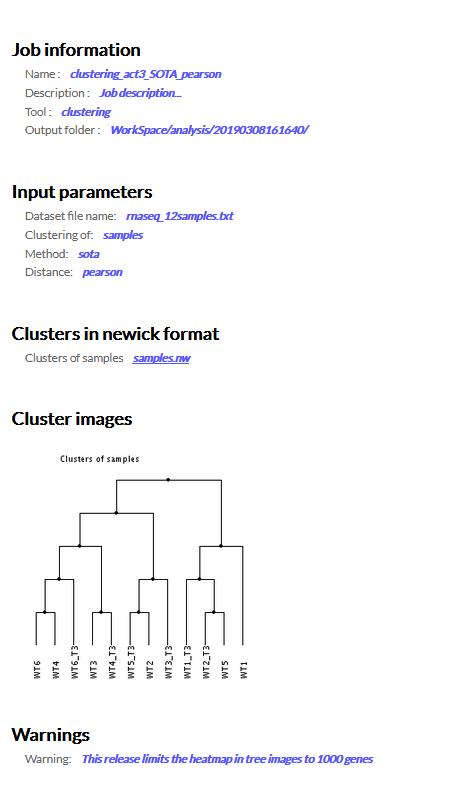












**C.2. Supervised classification**

**Predictors** are used to assign a new data (expression, proteins, metabolites…) to a specific class (e.g. diseased case or healthy control) based on a rule constructed with a previous dataset containing the classes among which we aim to discriminate. This dataset is usually known as the **training** set. The rationale under this strategy is the following: if the differences between the classes (our macroscopic observations, e.g. cancer versus healthy cases) is a consequence of certain differences an gene level, and these differences can be measured as differences in the level of gene expression, then it is (in theory) possible finding these gene expression differences and use them to assign the class membership for a new array. This is not always easy, but can be aimed. There are different mathematical methods and operative strategies that can be used for this purpose.

In [Babelomics](http://babelomics.bioinfo.cipf.es/), there is an unsupervised classification module to help in the process of building a “good predictor”. In this resource:

* We have implemented several widely accepted strategies so as this tool can build up simple, yet powerful predictors, along with a carefully designed cross-validation of the whole process (in order to avoid the widespread problem of “selection bias”).
* Babelomics allows combining several classification algorithms with different methods for gene selection.
* Main indicators to assess the quality of prediction: [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision), [MCC](https://en.wikipedia.org/wiki/Matthews_correlation_coefficient), [AUC](https://en.wikipedia.org/wiki/Receiver_operating_characteristic) and [RMSE](https://en.wikipedia.org/wiki/Root-mean-square_deviation).
* [More detailed information about methods](https://github.com/babelomics/babelomics/wiki/Class-prediction).

## Activities

We have prepared two activities to know how is possible the generation of predictors from Babelomics.

1. [Class prediction in acute leukemia](http://bioinfo.cipf.es/WODA/doku.php/supervised.class/exercises/ej1).
2. [Supervised classification for RNA-Seq data of Lung squamous cell carcinoma](http://bioinfo.cipf.es/WODA/doku.php/supervised.class/exercises/ej2).

Here you have more [detailed information](https://github.com/babelomics/babelomics/wiki/Tutorial-Class-prediction) about supervised classification module in Babelomics

# Activity 1. Class prediction in acute leukemia

In this example we are going to analyse a dataset from Golub et al. (1999). In that paper they were studying two different types of leukemia (acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) in order to detect differences between them. This dataset have 3051 genes and 38 arrays, 27 of them labeled as ALL and 11 of them as AML.

Using Class prediction we are going to build a predictor to try to distinguish between both classes. In the train file we can see 30 arrays, 21 ALL and 9 AML. The rest, 6 ALL and 2 AML, are in the test file for predicting.

You can find the dataset for this exercise in the following files:

* The first one is the file to train the predictor: [datatraingolub.txt](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/supervised.class/exercises/datatraingolub.txt).
* The second one will be used to predict the classes (test dataset): [datatestgolub.txt](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/supervised.class/exercises/datatestgolub.txt).

Principio del formulario

Final del formulario

## A. Training

* Train with KNN algorithm. Upload the datafile and select the variable TUMOR. In order to get the exercises fast select 5 repeats of 5-fold cross validation. In this exercise do not select any feature selection method.
* Repeat the exercise but select CFS feature selection method, which one works better? Why? how many genes were selected
* Now try with SVM algorithm with no feature selection method, which one performs better? SVM or KNN
* To finish you can try SVM with CFS feature selection method, how many features were selected? Why it matches KNN with CFS?
* Finally, which is the best combination? Why is SVM doing better along than with CFS?

Principio del formulario

Final del formulario

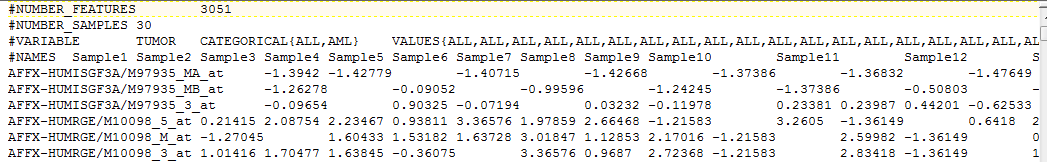
## B. Test

* Now we select the option Train and test and select datatraingolub and datatestgolub.
* We can select KNN without feaure method to speed up the exercise.
* In order to check the accuracy of prediction you can see the correct labels for the test file:

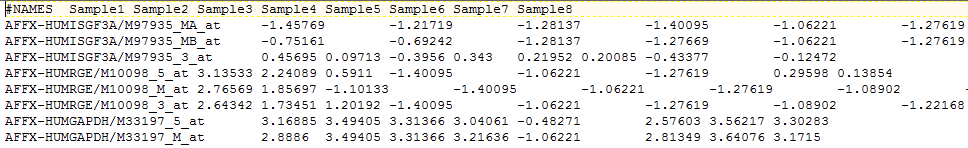
ALL ALL ALL ALL ALL ALL AML AML

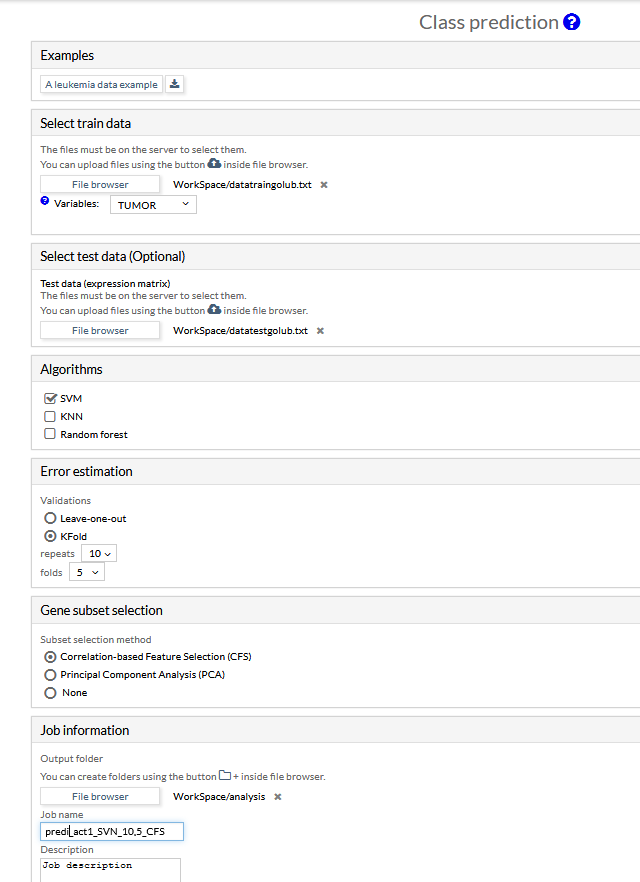
* Are the predictions right? Do you get the same results with SVM?

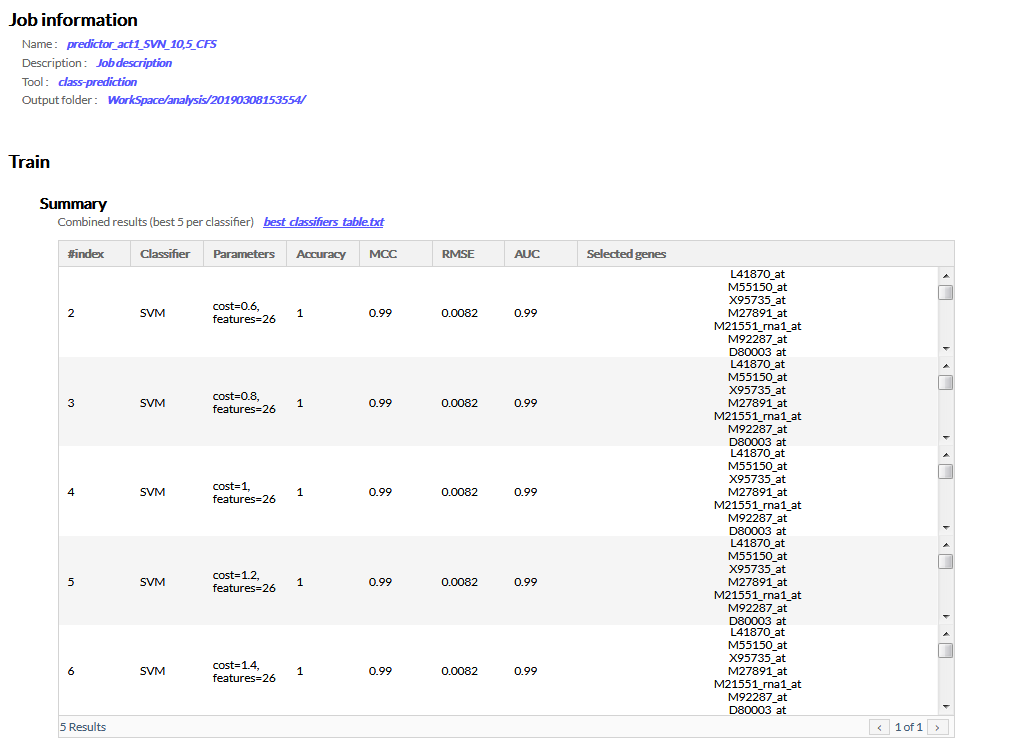
Train data:

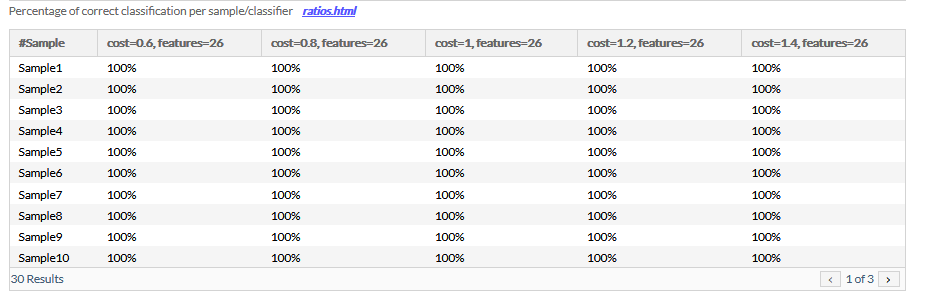


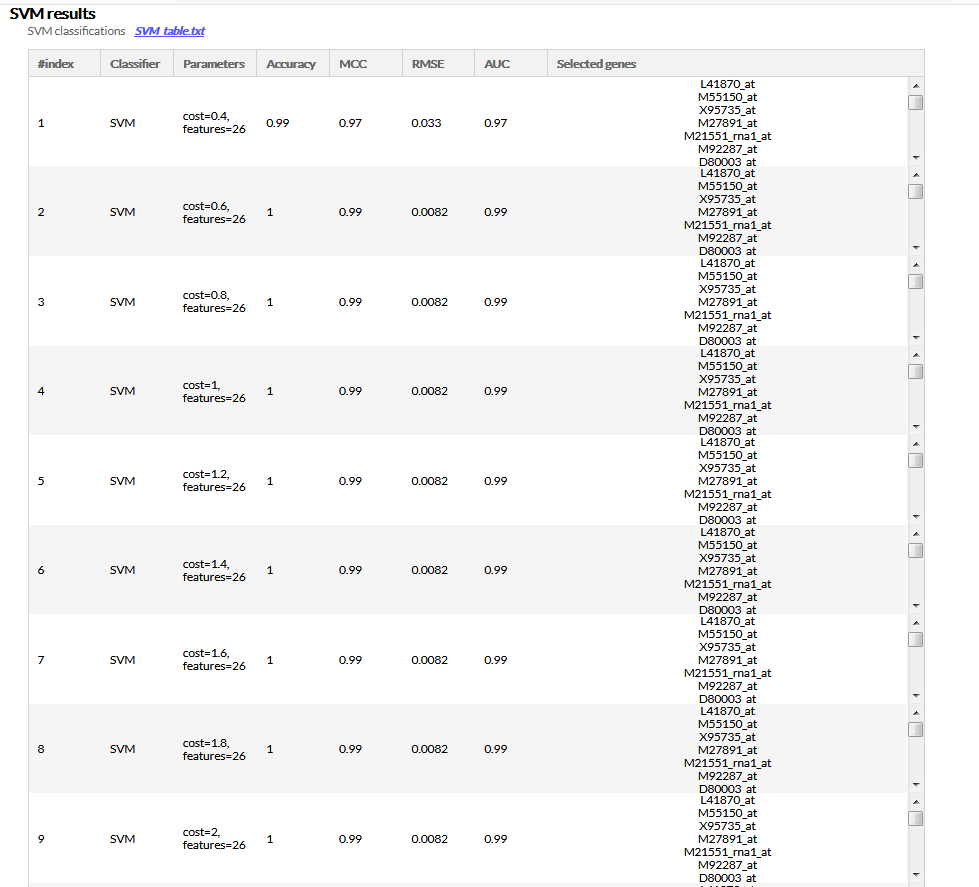
Test data

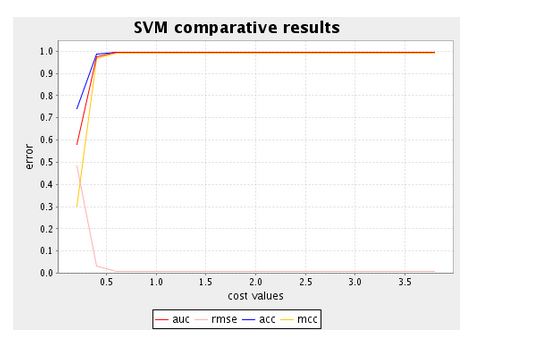


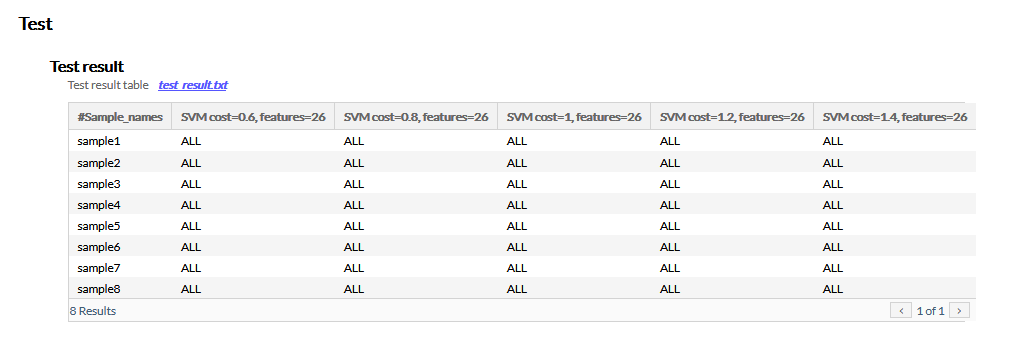












# Activity 2. Supervised classification for RNA-Seq data of Lung squamous cell carcinoma

Principio del formulario

Final del formulario

## Data description

RNA-Seq data of Lung squamous cell carcinoma (LUSC) samples taken from [The Cancer Genome Atlas (TCGA)](https://portal.gdc.cancer.gov/projects/TCGA-LUSC) data portal.

Principio del formulario

Final del formulario

## Goals

1. We want to train several classification models in [Babelomics](http://babelomics.bioinfo.cipf.es/).
2. After this step, we are evaluating the best way of classifying our data from a test dataset.

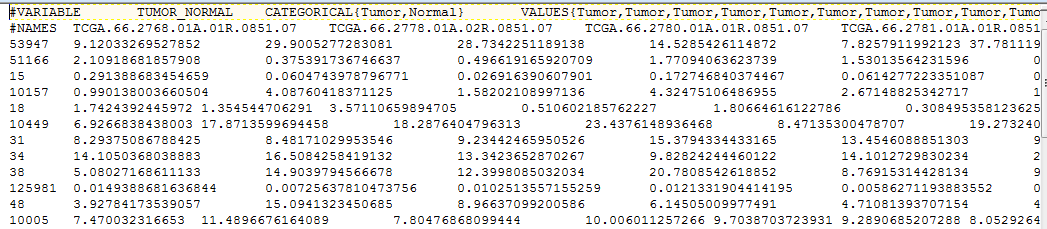
Principio del formulario

Final del formulario

## Work plan

1. Download [tca\_gene\_lusc\_train.txt](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/supervised.class/exercises/tcga_gene_lusc_train.txt). Contains 11 Normal and 150 Tumor samples.
2. Download [tca\_gene\_lusc\_test.txt](http://bioinfo.cipf.es/WODA/lib/exe/fetch.php/supervised.class/exercises/tcga_gene_lusc_test.txt). Contains 6 Normal and 75 Tumor samples.
3. Upload your files to Babelomics 5.0. Go to section Expression > Class Prediction
4. Try several classification strategies:
   * Select SVM, KNN and Random Forest
   * Select Leave-one-out for error estimation
   * Select Correlation-based Feature Selection (CFS)
5. Download test\_result.txt
   * Which supervised classification method(s) works better?
   * How many genes were used for the prediction?
   * Are the selected genes same for all methods?

Train data:



Test data:

