# Laboratory Of Biological Data Mining
# Project Proposal

Aliaksandr Siarohin, Mat. number 180402

## Problem:

Using genes expression data, find genes which are related to genes in flavonoidi pathway.

## Solution:

Let D be gene expression matrix. D – will be (n, m) – matrix, where n – number of genes and m – number of experiments.

We will take every gene in a pathway (or just target genes) and try to find genes which are related to them. For now let consider only one gene (We apply similar procedure for all of them). Let number of this gene be k.

Finding genes related to gene number k:

1. Create prediction problem from data as follows: $X = D^T_{(1,2,..k-1,k+1,..n)}$ (genes expression data for all genes except gene number k) and $Y = D^T_k$ (genes expression data just for gene number k) .
   The prediction problem now is given X predict Y. (We interpret every gene as feature an every experiment as train example)
2. Use one feature selection method from listed in http://scikit-learn.org/stable/modules/feature_selection.html (for example LassoSVM), in order to find we are the most relevant features in this prediction problem (which corresponds to the most relevant genes for predicting the expression of gene number k).
3. Try different values for filtering parameter (which is parameter that controls how much features we will discard). Choose target accuracy, for example x (maximum allowed error). Chose filtering parameter equals to value which leaves minimum number of features, but accuracy should exceed x. Visualize this step (Plot number number of discarded features against filtering parameter, as well as prediction error against filtering parameter).

## Result Evaluation:

1. Plot the statistics $\frac{NPathwayFeaturesRest}{NPathwayFeatures} - \frac{NFeaturesRest}{NFeatures}$ against filtering parameter. If it's grows this means we on the right way.
2. We can compare different feature selection methods if we take filtering parameter form step 3(Solution) and statistic from step 1(Result Evaluation).

## Final Result:

In order to obtain final result, we will use method that give best result in step 2 (Result Evaluation). We take filtering parameter from step 3 (Solution). The genes that will stay after applying this method will be the final result.