**State the objective of the project, the research to be carried out, and reasons for choosing the research topic**

Thank to the cooperation with the Institute of Bioorganic Chemistry of the Polish Academy of Sciences and participation in the Acute Myeloid Leukemia project I realized that very important step during investigation is to predict from which class (e.g. AML or ALL) samples (e.g. patient) come from. During our research it was noticed that building the classifiers can not be based on too large number of genes since it leads to overfitting of the classifier. For this reason, we start to think about how to reduce the list of significant genes. It turned out that taking into account the correlation between the expression values have a significant impact on the quality of the prediction.

Therefore, the main objective of the project is the construction of a statistical tests by which the researcher will be able to find the interesting genes that would give the best efficiency in the prediction of samples to particular classes (e.g. healthy - sick) based on the gene expression values and the correlation between them. In addition, in the classification process we want to avoid the selection of groups of highly correlated genes, because such genes do not bring too much additional information in relation to the subsequent classification. It also means that they generate similar errors in the use of cross-validation method. The simplest significance analysis, based on assumptions that the lowest correlation coefficient between significant genes is desired has however not been used yet, to the best of our knowledge. The second objective consists of the comparison of the effectiveness of existing methods used in gene selection problems with the proposed novel statistical tests. On the basis of preliminary studies, we can conclude that taking into account the correlation effect between the values of gene expression has a significant impact on the discriminant analysis. The main hypothesis of the research project are as follows: selection method in which the weak correlation between significant gene expression measurements are taken into account can cause a higher performance of the discriminat analysis and there exist correlation coefficient which used during selection process have an impact on increasing efficiency of prediction samples to the classes. During examination the correlation between genes it will be taken into account different factors widely use in the construction of co-expression networks such as: the mutual information, Spearman rank correlation, Euclidean distance, the partial correlation, the Gini coefficient and the Kendall ratio.

In the SAMaRco project microarray and RNA-seq data simulation will be conducted in order to create data with specific correlation structures and with known classes. During the project statistical model used in the subsequent analysis of microarray and RNA-seq data will be proposed. This model will be based on network adjacency matrix. The proposed construction of a statistical tests will base on: the expression values and the values of the correlation between gene expression levels. In the procedure more impact will have genes, that have the greatest differences in expression levels between classes and at the same time the lowest coefficients of correlation between the levels of expression. For the validation of the novel statistics we will use following classifiers: naive Bayes, neural network, nearest neighbor, support vector machine and random forest. Leave One Out Cross and k-fold Validation will be used to calculate prediction errors. At the end, new statistical tests will be compared with the existing statistics methods used for considered problem. In this comparison it will be taken into account both: traditional tests as well as the ones that take into account the correlation structure. The analysis will be carried out using the ROC curve, AUC values and diagnostic plots. All analysis will be carried out using the free software R and BioConductor platforms dedicated to statistical analysis dedicated to genetic data.