*Faculty of Computers and Artificial Intelligence*

*Bioinformatics Program*

*2021/2022*

*Biology Statistics Course*

Assignment 3

Statistical Feature Analysis

**This assignment is to be solved in teams of 2.**

**Deadline is 4/1/2022**

The objective of this assignment is to apply data processing and statistical methods learned to perform feature selection with different methods and compare them to select best features.

Some methods you took that can be used for feature selection:

* Correlation coefficient
* Hypothesis testing : Student T-test

# The dataset

## Breast Cancer Proteomes Dataset

Breast Cancer is a multifactorial disease that forms in the cells of the breast. Breast cancer can occur in both men and women, but it's far more common in women.

## Dataset Description

This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). It contains expression values for ~12.000 proteins for each sample, with missing values present when a given protein could not be quantified in a given sample.

This dataset consist of 3 tables:

1. **77\_cancer\_proteomes\_CPTAC\_itraq.csv**

This table includes protein expression values and metadata on 12553 genes from 80 breast cancer patients and 3 healthy individuals.

* RefSeq*accession*number: RefSeq protein ID (each protein has a unique  
  ID in a RefSeq database)
* gene\_symbol: a symbol unique to each gene (every protein is encoded  
  by some gene)
* gene\_name: a full name of that gene  
  Remaining columns: log2 iTRAQ ratios for each sample (protein  
  expression data, most important), three last columns are from healthy  
  individuals

1. **clinical\_data\_breast\_cancer.csv**

This table contains clinical data and various breast cancer classifications from 105 breast cancer patients

* First column "Complete TCGA ID" is used to match the sample IDs in the main cancer proteomes file
* Other columns contain data about the cancer classification of a given sample using different methods

1. **PAM50\_proteins.csv**

This table contains the list of genes and proteins used by the PAM50 classification system. The column RefSeqProteinID contains the protein IDs that can be matched with the IDs in the main protein expression data set.

# Part1: Data preprocessing (1.5 grades)

* 1. (1) Subset the columns of proteomes table. As you learned about genomic datasets you will find:
* The columns in proteomes table represent patients samples
* The rows in clinical data table also represent patients samples

You will find that number of columns don’t match so you need to subset the data in both proteomes table and clinical data table to keep the data of patient samples that exist in both tables. (Like inner join in DBs)

**Note:** you will find that the IDs in the two tables don’t have the same format so you will need to find a way to match the formats.

* 1. (0.5) Subset the rows of proteomes table by removing rows with missing values

# Part2: Feature Selection with correlation coefficient (2.5 grades)

2.1 (1) Apply Pearson correlation coefficient between the each protein (each protein is considered as feature) from the protein table and the ***HER2 Final Status*** column in the clinical data table

2.2 (1) Order the result from the highest correlation (regardless positive or negative) to the lowest correlation

2.3 (0.5) Based on the results from 2.2, set a suitable threshold and filter the list by the proteins with correlation value greater than this threshold.

# Part3: Feature selection by hypothesis testing

Hypothesis testing is a methodology in statistics that allows for applying statistical testing for individual features. Feature selection via hypothesis testing will attempt to select only the best features from a dataset. The hypothesis can be defined as:

**True or False: This feature has no relevance to the response variable**

We want to test this hypothesis for every feature and decide whether the features hold some significance in the prediction of the response. The significance level can be tested by the p-value.

## Feature selection with T-Test (4 grades)

In classification problem, t-Statistics helps us to evaluate whether the values of a particular feature for class A is significantly different from values of the same feature for class B. If this holds, then the feature can helps us to better differentiate our data.

We need to decide for each protein if it has impact on “**HER2 Final Status**” value. Steps (for each protein):

* 1. (1) Split into two groups of values (group for samples with +ve HER2 and group for samples with –ve HER2 Final Status)
  2. (1) Use t-test to check whether the two groups are significantly different or not
  3. (1) After calculating the values of t-Statistic for each feature, sort these values in descending order in order to select the most affective features given a selected threshold.
  4. (1) Compare between the results of part 2 and part 3, did both methods resulted in the same feature set