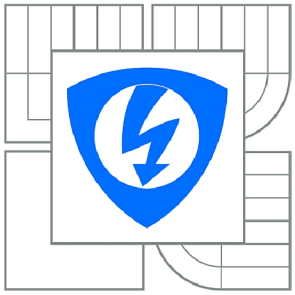
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

MOLECULAR SIGNATURE AS OPTIMA OF MULTI-OBJECTIVE FUNCTION WITH APPLICATIONS TO PREDICTION IN ONCOGENOMICS

MOLEKULÁRNÍ SIGNATURA JAKO OPTIMÁLNÍ MULTI-OBJEKTIVNÍ FUNKCE S APLIKACÍ V PREDIKCI V ONKOGENOMICE

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. ZUZANA ALIGEROVÁ

AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. IVO PROVAZNÍK, Ph.D.

SUPERVISOR

BRNO 2015

BRNO UNIVERSITY
OF TECHNOLOGY

Faculty of Electrical Engineering and
Communication

Department of Biomedical Engineering

Diploma thesis

Master's study field
Biomedical and Ecological Engineering

Student: Bc. Zuzana Aligerová
ID: 136460
Year of study: 2
Academic year: 2014/15

TITLE OF THESIS:

Molecular Signature as Optima of Multi-Objective Function with Applications to Prediction in Oncogenomics

INSTRUCTION:

1) Study classical discrete optimization methods, mainly dynamic programming and greedy optimization, and their limits. Study the basis of meta-heuristic methods for multi-objective optimization. 2) Focus on the basis of oncogenomics, mainly on the general problem of prediction in transcriptomic datasets (RNA microarrays), statistical validation of the predictors, main methods of molecular signature computation, and the application of meta-heuristic methods for molecular signature computation. 3) Implement selected methods of gene selection and test them on real datasets. 4) Implement at least one method of molecular signature computation by meta-heuristic optimization. 5) Design and statistically validate predictors in oncogenomics (various organs and various datasets, including Next Generation Sequencing data). 6) Evaluate and discuss results.

REFERENCE:

[1] Richard Neapolitan, Kumarss Naimipour: Foundations Of Algorithms Using Java (or C++); 3rd edition; Jones and Bartlett Publishers; 2004.
[2] Francisco Azuaje: Bioinformatics and Biomarker Discovery (“Omic” Data Analysis for Personalized Medicine); John Wiley & Sons, 2011.

Assigment deadline: 9. 2. 2015 ***Submission deadline:*** 30.7.2015

Head of thesis: prof. Ing. Ivo Provazník, Ph.D.
Consultant:

prof. Ing. Ivo Provazník, Ph.D.
Subject Council chairman

WARNING:

Abstrakt

Náplní této práce je teoretický úvod a následné praktické zpracování tématu Molekulární signatura jako optimální multi-objektivní funkce s aplikací v predikci v onkogenomice. Úvodní kapitoly jsou zaměřeny na téma rakovina, zejména pak rakovina prsu a její podtyp triple negativní rakovinu prsu. Následuje literární přehled z oblasti optimalizačních metod, zejména se zaměřením na metaheuristické metody a problematiku strojového učení. Část se odkazuje na onkogenomiku a principy microarray a také na statistiku a s důrazem na výpočet p-hodnoty a bimodálního indexu.

Praktická část je pak zaměřena na konkrétní průběh výzkumu a nalezené závěry, vedoucí k dalším krokům výzkumu. Implementace vybraných metod byla provedena v programech Matlab a R, s využitím dalších programovacích jazyků a to konkrétně programů Java a Python.

Klíčová slova:

Optimalizační metody, Strojové učení, triple negativní karcinom prsu, Lineární diskriminantní analýza, onkogenomika, microarray, RNA, gen, molekulární podpis, p-value, index bimodality, Matlab, R

Abstract

Content of this work is theoretical introduction and follow-up practical processing of topic Molecular signature as optima of multi-objective function with applications to prediction in oncogenomics. Opening chapters are targeted on topic of cancer, mainly on breast cancer and its subtype Triple Negative Breast Cancer. Succeeds the literature review of optimization methods, mainly on meta-heuristic methods for multi-objective optimization and problematic of machine learning. Part is focused on the oncogenomics and on the principal of microarray and also to statistics methods with emphasis on the calculation of p-value and Bimodality Index.

Practical part of work consists from concrete research and conclusions lead to next steps of research. Implementation of selected methods was realised in Matlab and R, with use of other programming languages Java and Python.

Keywords:

Optimization methods, Machine learning, triple-negative breast cancer, Linear discriminant analysis, oncogenomics, microarrays, RNA, gene, molecular signature, p-value, Bimodality index, Matlab, R

ALIGEROVÁ, Z. *Molecular Signature as Optima of Multi-Objective Function, with Applications to Prediction in Oncogenomics*, Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 2015. 63 p., Adviser prof. Ing. Ivo Provazník, Ph.D., Ing. René Natowicz

Declaration

I declare that I have written my final project on the theme of "Molecular Signature as Optima of Multi-Objective Function, with Applications to Prediction in Oncogenomics" independently, under the guidance of the project superviser and using the technical literature and other sources of information which are quoted in the project and detailed in the list of literature at the end of the project.

As the author of the project I furthermore declare that, as regards the creation of this final project, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal and/or ownership rights and I am fully aware of the consequences in the case of breaking Regulation § 11 and the following of the Copyright Act No 121/2000 Sb., and of the rights related to intellectual property right and changes in some Acts (intellectual property Act) and formulated in later regulations, inclusive of the possible consequences resulting from the provisions of Criminal Act No 40/2009 Sb., Section 2, Head VI, Part 4.

In Brno ..... Author .....

Acknowledgement

I would like to thank to my supervisor, Pr. René Natowicz, for his patient, all advices and help during time when I was part of the project qnd for leading my work. Further, I thank to Pr. Thiago Souza, from university of Brazil, for all his help and support and finally to Pr. Patrick Siarry, Pr. Fabien Reyal and prof. Ing. Ivo Provazník, Ph.D. for their support.

In Brno ..... Author .....

Content

List of figures

Introduction

The problematic of the cancer generally is one of the most discussed problems all around the world. The breast cancer is the biggest killer for the women with a lot of subtypes. Thanks to the gene expression measurement technologies the breast tumors could be categorized into these molecular subtypes that brought the development of symptoms and treatments which are still not completely clarify in any cases. One of the most aggressive and recurring type is basal-like or let say triple negative breast cancer subtype. Also this type of breast cancer does not react on the traditional types of chemotherapy and its prediction is difficult. The basal-like prognosis is bad and in this particular case there is no single gene whose distribution of expression levels are significantly different in the class of patients who are responders to chemotherapies and to who are not. It is why a lot of studies are focused on the ways of robust prediction of the outcome of chemotherapy treatment. Genomic predictors of the responses to chemotherapy take often as input the expression levels of a subset of genes (a molecular signature) and combine them into a classification function that allocates the patient cases to the pCR/noPCR phenotype. It would state a sub-classification of the basal-like molecular class, these opening research tracks to designing new treatments dedicated to the basal-like noPCR patient cases. These solutions and conclusions are very important for clinicians and for searching of right cure.

This research is focused on the application of optimization methods to find required prediction in chemotherapy for this breast cancer subtype.

The first chapters take a closer look on cancer and mainly on the breast cancer and its subtype Triple Negative Breast cancer which is hard predictable and the most aggressive type of cancer and the main subject of this study. Next ones give us the review about problematic of oncogenomics with description of RNA and function of the microarray techniques and their use for clinical and statistical research and statistical validation. Theoretical preparation also takes a closer look to optimization and metaheuristic methods with a more detailed analysis of machine learning problematic.

And the end of theory is devoted to statistical methods as Mann-Wilcoxon-Whitney test and Bimodality index.

The last chapters contain from the research of gene expression analysis and prediction of chemotherapy response in breast cancer realized on ESIEE Paris. In research are used Machine learning method of Linear discriminant analysis and Wilcoxon Mann Whitney test, which is statistic method for p-value finding. Because of importance of robustness later the research digress from supervised analysis to use method of unsupervised analysis, specifically Bimodal Index.

1 Cancer

Cancer is recognized as a heterogeneous disease with distinct subtypes and outcomes that can be predicted by a limited number of biomarkers. Cancer is highly complex disease which can encompass multiple genomic alterations, including point mutations, translocations, gene amplifications, epigenetic modifications, deletions, aberrant splicing, and altered gene expression. These changes may be inherited or somatically acquired during progression from a normal to a cancerous cell.

The reason of the beginning of cancer, also called malignancy, is an abnormal growth of cells so it is genetic disease caused by accumulation of mutation to DNA leading to unrestrained cell proliferation and neoplasm formation. This DNA is not able to repair itself or cause the cell die, cells start to grow and divide uncontrollably and change to the cancer cell. Cancer cells multiply and display normal cells, as the tumour is larger it develops his own blood supply. Since cancer cells do not stick together as normal cells they can change the place and enter the blood vessels or lymphatic system near the tumour to travel to locations in the body and form additional tumours. It is referred as metastatic or advanced cancer.[10], [18]

1.1. World disease

The global cancer burden has changed dramatically over time. And still it is one of the most growing and important field in biomedicine.

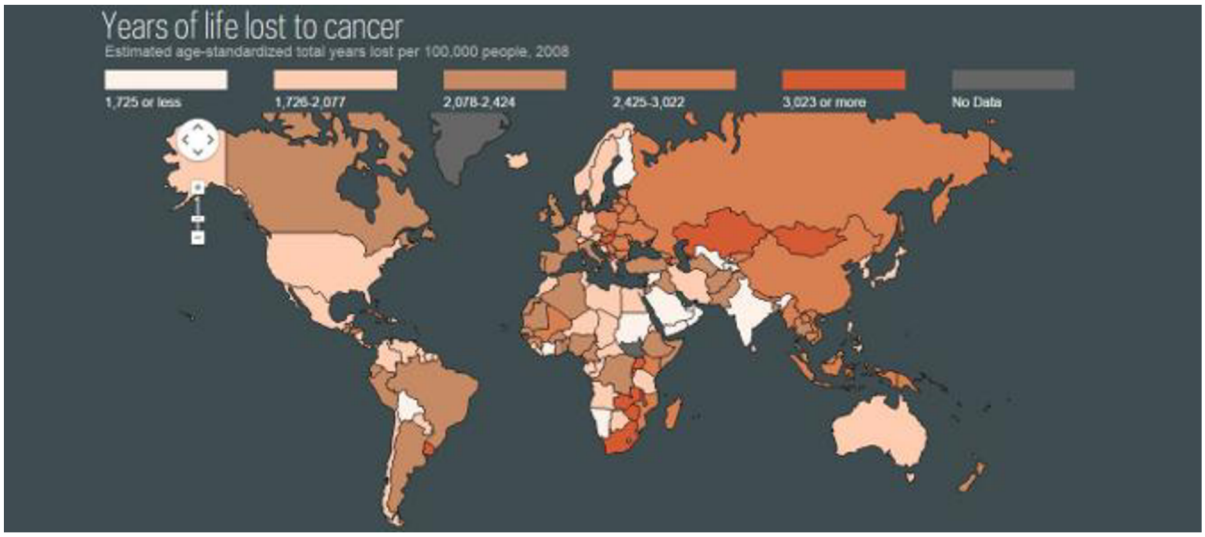


Figure 1 - Global cancer map [9]

In the figure 1 we can see decomposition of sickness on the world. For example the countries with rank low on the human development index (Sub-Saharan Africa) suffer from relatively high rates of cervical cancer. More developed countries (United States, Canada) have been able to bring down their cervical cancer rates though more Pap testing, there are also other types of cancer which are problematic for Western

countries. For example lung cancer is higher in countries which were historically depending on tobacco farming and production. [9]

1.2. Cancer types

There are more than one hundred types of cancer, including breast cancer, skin cancer, lung cancer, colon cancer, prostate cancer and lymphoma. Symptoms depend every single type. Gene-expression biomarkers have enabled the identification of sub-classes of cancers and prognostic signatures in breast and lung cancers.

For every specific type of cancer there is prevention and then tests and, if suspicion, how to found and cure relevant type (X-rays, medical and surgical history, smoking and work history, CT, MRI, biopsy, bronchoscope, sputum testing). [10], [16]

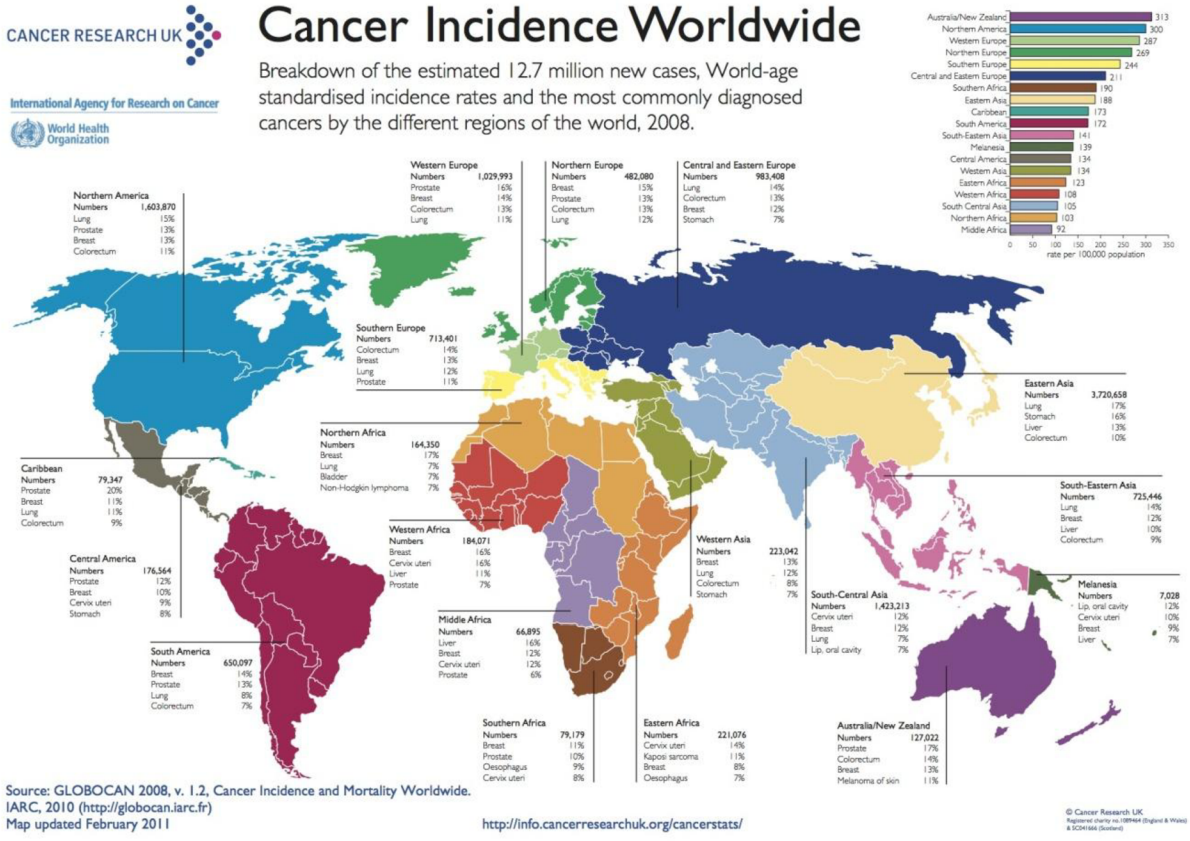


Figure 2 - Map of worldwide cancer incidence, updated 2011 [46]

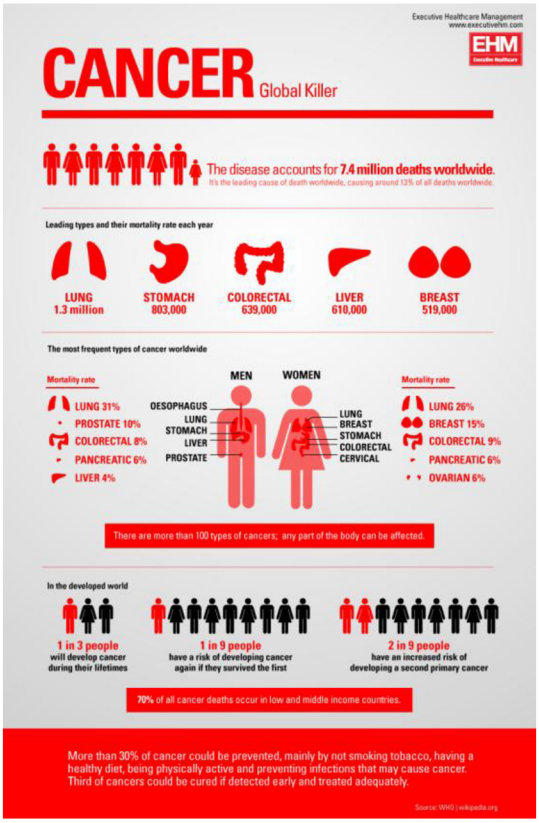


Figure 3 - the most killing type of cancers [published by Executive healthcare]

Around the world the most dangerous and killing cancers are lung, prostate and breast cancers. Figure 2 show the worldwide incidence of cancer and figure 3 shows the main cancer killers and their representation.

The aim of this research is to find new ways of prediction in breast cancer, namely triple negative subtype, which is difficult to predict.

1.3. Cancer Treatment

For the cancer therapy we can use more types of treatment like surgery, chemotherapy, radiation therapy, hormone therapy, targeted therapy, bone-directed therapy and medicaments. The most often use type of treatment next to surgery is probably chemotherapy. Treatment can include chemotherapy, radiation and surgery. Also for some types of cancer, e.g. prostate, we can use cryotherapy and hormonal therapy. [10]

1.4. Chemotherapy

Chemotherapy is a type of treatment that includes a drug or combination of drugs to treat cancer. The goal of chemo is to cure the cancer, keep the cancer from spreading, stop or slow the growth of cancer cells, kill cancer cells that can spread to the rest of body and relieve symptoms caused by disease. It is considered as a systemic therapy, this means that it can affect all body and work throughout it.

Chemotherapy is dividing to the two types called adjuvant and neo-adjuvant chemotherapy. Neo-Adjuvant chemotherapy is the chemotherapy used before surgery or radiation therapy to shrink the tumour. This type has a pathologic complete response and its effect is unclear and the opinions of clinicians are different. On the other side adjuvant chemotherapy is used after surgery and should be effective for preventing disease recurrence.

There is the increasing number of the chemotherapy options, more than one hundred chemo drugs are used in many combinations with different side-effects, i.e. however, chemotherapy not even affects rapidly growing cancer cells but they can also affect healthy cells that grow rapidly like white blood cells, red blood cells and platelets, also hair follicles have cells that can affect by chemotherapy and it is leading to hair loss, also called alopecia, cells lining in stomach can be affected too. This cause vomiting and diarrhoea and may be associated with nausea (for example Platinum-based drugs, which caused nausea, vomiting, kidney and nerve damage or other side eribulin which caused fewer, damage of gastrointestinal sector and hair loss and we have more other examples). Other side effects can be trouble with memory, mouth sores or fatigue. Also it is important to be careful because of infections, since immunity is weak during and after chemotherapy. But nowadays lot of these side effects can be managed.

Because of use of the lot of combinations and drugs for chemotherapy there is decreasing number of resistance cancers to this type of treatment or possibilities of local chemotherapy treatments.

Because of these aspects it is important to find good diagnose and tailor chemotherapy according to individual patient and tumour variables. [10], [46]

2 Breast cancer

Breast cancer (BC) is the most common cancer for women and the second leading cause of cancer deaths after lung cancer. Two-third of women with breast cancer is over 50, and the most of the rest are between 39 and 49. It is a reason why there is mandatory testing in Czech Republic up to 40. Only 1% of breast cancer is diagnosed to men. Studies show that there is the connection between some types of BC and ethnicity.

As in all forms of cancer, the abnormal tissue that makes up breast cancer is the patient´s own cells that have multiplied uncontrollably. BC develops in the breast tissue, primarily in the milk ducts (ductal carcinoma, the most common type) or glands (lobular carcinoma). It usually begins with the formation of a small, confined tumour (lump) or as calcium deposits (micro calcifications) and then spreads though channels within the breast to the lymph nodes or to the other organs through the blood stream. The tumour than may grow and invade tissue around.

The symptoms connected with the breast cancer are at figure 4.

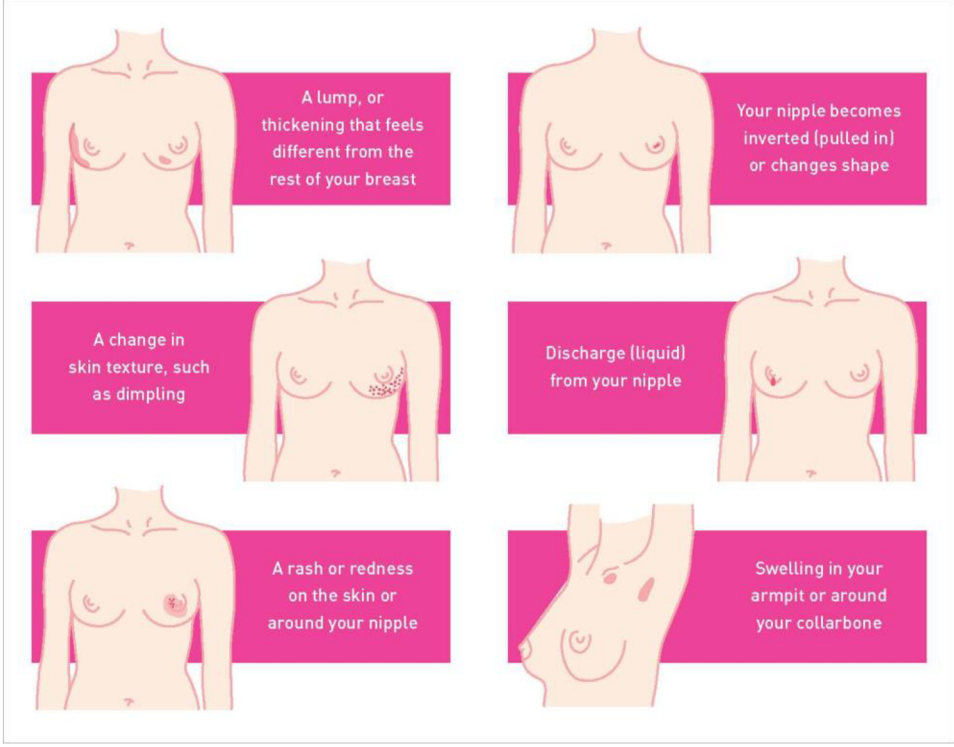


Figure 4 - Breast cancer symptoms [48]

BC subtypes can be separated to different types based on the way the cancer cells looks under the microscope. Most of them are carcinomas, a type that starts in cells, i.e. epithelial cells, that line organs and tissues. Namely this type of carcinoma for BC is called adenocarcinoma and starts in glandular tissue. In some cases a breast tumour can be mixture of invasive and in situ cancer or combination of different types. Cancer in situ is considered non-invasive or pre-invasive, which means that the cells

have not spread outside the breast. However, some cases of the Ductal carcinoma in situ can go on to become invasive, but for now there is no study how to know which cases will be invasive and which ones will not. There are two types of the invasive carcinoma, depends on the place of start, ductal and lobular carcinoma. Next to these mostly common types of cancer we can find the less common types as inflammatory BC, paget disease of the nipple, phyllodes tumour, angiosarcoma and specially invasive breast carcinoma as adenoid cystic, papillary carcinoma, mixed carcinoma, micropapillary carcinoma and etc. [10]

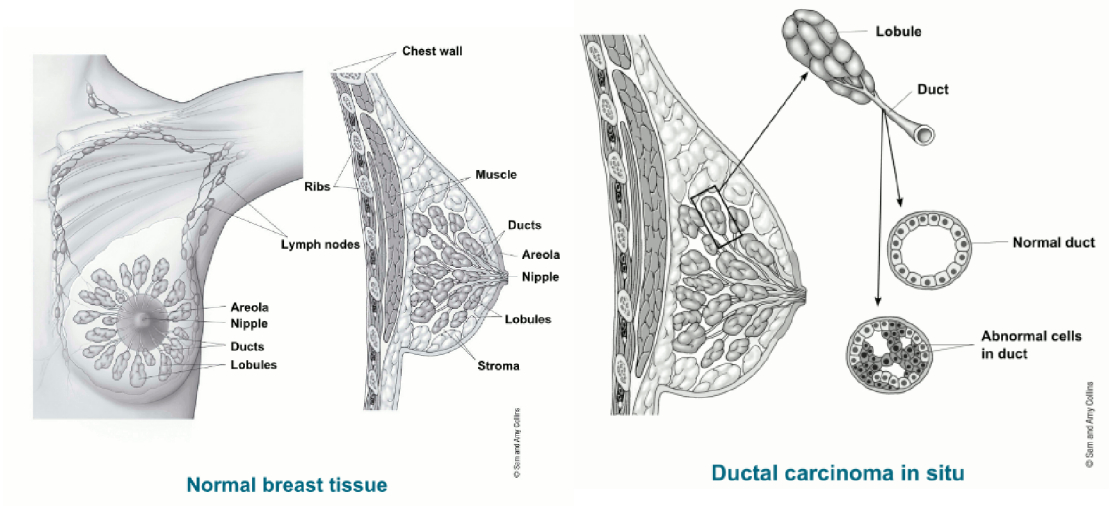


Figure 5 - Difference of the BC cells for Normal breast tissue and breast tissue with dinase - Ductul carcinoma in situ [10]

BC is in other way classified based on proteins on or in the cancer cells.
Test to classify breast cancer are estrogen receptor (ER), progesterone receptor (PR) and HER2/neu testing. Receptors are proteins in or on certain cells that can attached hormones and especially these two hormones which normally help the breast cancer cells to growth. BC types contain from one or both of these receptors. BCs with ERs are called ER-positive and progesterone receptors cancers are called PR-positive.
Tumours with increased levels of HER2 are referred as HER2-positive and in this case they contain from a growth-promoting protein (HER2). From these can be BCs classified in way based on hormone receptors and HER2 status as hormone receptor-positive, hormone receptor-negative, HER2 positive, HER2 negative and triple negative or triple positive. To know the group of classification is important because of differences between reactions on the treatment. [10]

The last form of classification is the classification based on gene expression. This classification based on molecular features, divides BCs to five molecular subtypes that are presently established, these are Luminal A, Luminal B, HER2-positive, basal-

like (BLBC) and Claudin low, each with significantly molecular heterogeneity and different subset of genetic and epigenetic abnormalities. [30]

The Luminal types are ER-positive. The gene expression patterns are similar to normal cells that line the breast ducts and glands. Luminal A is slow growing and low grade subtype of cancer with the best prognosis. Luminal B is more aggressive and growth faster compared Luminal A and their outlook is not good. HER2 have extra copies of the HER2 gene and have tend to growth more quickly and have a worse prognosis than previous Luminal cancers, even these cancers often can be treated successfully thanks to chemotherapy therapies targeted on HER2.

Basal type lack estrogen or progesterone receptors and have normal amount of HER2, this cancer types are common for women with BRCA1 gene mutations and are also more common for young Afro-American women. These are high-grade cancers which grow quickly and have poor outlook. However, hormone therapy and anti-HER2 therapies (Trastuzumab, Lapatinid) are not effective against these cancer types. Chemotherapy can be helpful, but also not in every case. The less known type of breast cancer is Claudin low. In the papers of A. Prat and Ch. M. Perou is the one of the first mention about Claudin Low subtype. This subtype is related close to basal-like subtype, how is obvious from papers, first work of Prat and Perou shows that in their first opinion was Claudin Low subtype of the basal-like but in second study few years later they elucidated it like isolated subtype with similar features like basal-like. It is possible to use this subtype for compare of predictions. Even that this testing, called PAM50 is available, it is still not clear that it is more helpful in guiding treatment than previous test based on protein classification. [10], [26], [27]

All these classifications are the subjects of research and can be use to find the way how to get relevant data about breast cancer.

Studies are still lack a complete picture about biological heterogeneity of BC with respect to molecular alternations, treatment sensitivity and cellular composition. Moreover, this complexity is not totally reflected by the main clinical parameters,

e.g. age, node status, tumour size, ethnic, histological grade and etc. and pathological markers, i.e. estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2, all of which are normally used in the clinic for assessment of diagnosis and to select treatment. [26]

How was written in previous paragraphs the causes of breast cancer are unclear. Next to overweight, alcohol use, menopause, and hormone therapy are the most significant factors advanced age and family history. Logical risk increases little bit for woman who has certain benign, i.e. that tumours that are not cancerous, breast lumps and significantly for that one who had previously other type of cancer (endometrial, ovarian or colon). Women whose have female ascendant with cancer in family are two

to three times more prone to disease, but it does not mean that the disease will intervene them, around 85% women have BC coming from family history.

BRCA1 and BRCA2 are two genes responsible for this familial cancer. These genes predispose a woman to have the breast cancer, or also to ovarian cancer and are associated also with pancreas cancer and melanoma (BRCA2). Also other genes (PTEN, ATM, TP53, CHEK2) can increase the risk of breast cancer generally but this risk is lower and there can become different lists of top mutated genes from different studies. The group of mutated genes is significantly for every particular breast cancer class, for example for basal-like and triple negative gene breast cancer is called TP53, which is also the gene of ovarian cancer, this shows the connection between these two types of cancer. Lot of studies still are fastening on these mutated genes which can show us strictly on the disease and it can give material for one other article. [30]

Other studies are going around in the difference between risk for other ethnicity like African-American women, Europeans and Caucasians. There is the opinion that also hormone Estrogen affects the incidence of the breast cancer and some discussions are about effect of birth control pills. The conclusion is that there is lot of reasons and causes for BC, so the list can be long and still there are lot of theories and researches around this problematic.

Same like other cancer types also the breast cancer have different types which grow and spread at different rates, some need years while others grow and spread quickly. Some lumps are benign, means not cancerous, however these can be premalignant. [10]

Breast cancer is treatable if detected early and classification of the further cancer subtypes can throw light on the understanding of disease and better cure of cancer.

2.1. Triple-Negative Breast Cancer

All the breast cancers are characterized by using three biological factors ER, PR and HER2 expression status. However, Triple-Negative Breast Cancer (TNBC) is the bad responder to all of these three factors unlike other types of BC because of absence of the therapeutically targetable hormone receptors and HER2 protein over expression.

For this cancer subtype is also poor overall prognosis and there is no predictive biomarker of response or survival to allow tailored therapy for patients with disease. It means that Basal-like BC subtype comprises the majority of Triple Negative Breast Cancer, because of the lack ER and PR and normal amount of HER2. Unfortunately, other 20-30% of cases fall into other types. By this fact, that significant molecular heterogeneity even exists, i.e. BLBC has become more known as TNBC, but however, not all TNBCs can be identified as BLBC by gene expression. There are the various differences between cohorts in the proportion of BLBC within TNBC, it can be explain

by false negative immunohisto chemical results, otherwise the significant biological heterogeneity exists and the aim is to elucidate the clinical heterogeneity of TNBC as defined by three biomarkers for identifying the known intrinsic subtypes. This can be used for improving clinical outcomes and adapted therapy will require further stratification by biologic subtype.

All these aspects leading to that TNBC is the most aggressive BC from the all known types. It is spreading very fast and aggressively and mainly extends to lung and brain. Also it is the most recurring breast cancer after treatment, the five year survival rates tend be lower than for other types. Prediction of this cancer type is very difficult. Thanks to this specification is this breast cancer subtype responsible for a huge number of the breast cancer deaths, because these factors also influence prognosis and the available treatment options, which are in this case limited to chemotherapy. In spite of that fact TNBC has typically higher rates of chemo sensitivity compared with hormone receptor-positive BCs. [10], [28]

3 Basis of oncogenomics

Oncogenomics is a relatively new sub-field of genomics that applies "high trough put technologies" to characterize genes associated with cancer. The goal of oncogenomics is to identify new oncogenes or tumour suppressor genes. The steps and main goals of oncogenomics are shown at the figure 6. [17]

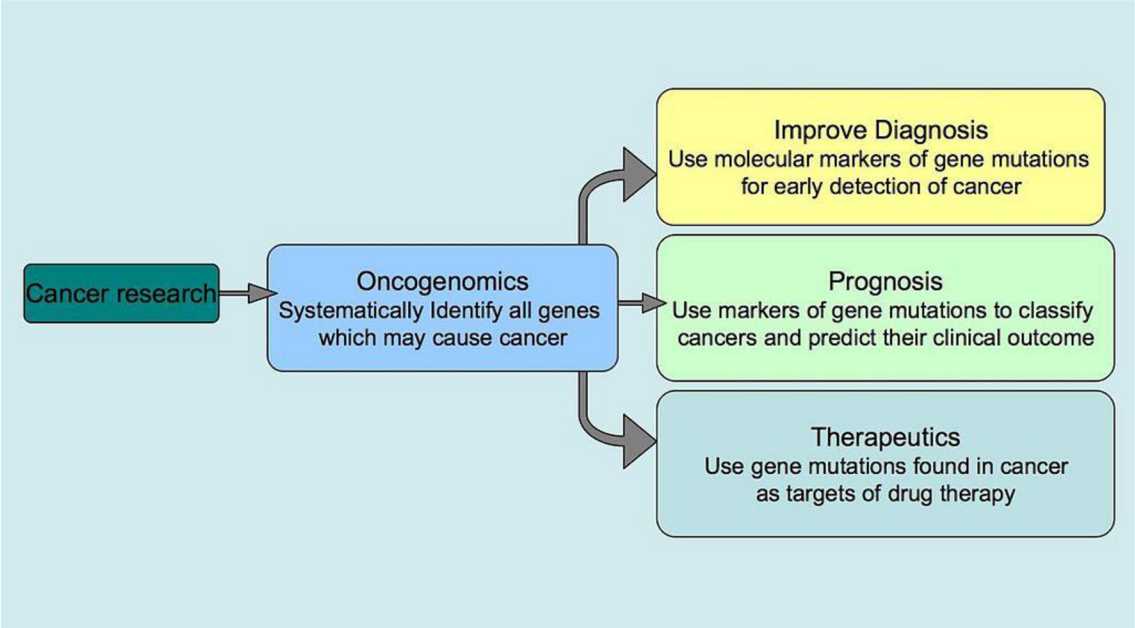


Figure 6 - Goals in oncogenomics [17]

3.1. Datasets for oncogenomic research

Biological databases dedicated to cancer data and oncogenomic research are occasionally available to cancer researchers as resources which have banked oncogenomic research data. There are some public available datasets for use in the research, e.g. Cancer Genome project (somatic intragenic mutations in cancer), Cancer Genome Anatomy Project (information of research on cancer genome, transcriptome and proteome), Progenetix (cytogenetic tumour data), Oncomine, IonOGen and RTCGD. Other datasets comes from special laboratories around whole world. There are just few professionals with deep specialization for getting data from the cancer tumour samples. [17]

3.2. Prediction in transcriptomic datasets

Biomarkers and gene expression data analysis can be guided to following topics as biomedical findings and clinical applications, statistical and data mining methodologies applied strengths and limitations. Changes in gene expression can be measured by different types of techniques ranging from smaller to large-scale approaches, and different in terms of their reliability and genome coverage. Traditional

gene expression analysis for cancer biomarker discovery has comprised the profiling of in vitro or in vivo tissue from tumours. [16]

4 RNA

Ribonucleic acid (RNA) is one of the three major biological macromoleculesm other two are DNA and protein, that are essential for all known forms of life.

It is a polymeric molecule made up of one or more nucleotides. Each nucleotide is made up of a base (adenine, cytosine, guanine and uracil), ribose sugar and a phosphate. Instead the DNA, RNA has a hydroxyl (-OH) group. RNA is typically found in a single-stranded form. The lack of a paired strand allows RNA to fold into complex, three-dimensional structures. RNA is important for protein synthesis and gene regulation. Primarily it is synthesized from DNA by an enzyme known as RNA polymerase during a process called transcription. The new RNA sequences are complementary to their DNA template, rather than being identical copies of the template. RNA is then translated into proteins by structures called ribosome.

There are three types of RNA involved in the translation process: messenger RNA (mRNA), transfer (tRNA) and ribosomal RNA (rRNA).

A "Central Dogma" of molecular biology tells that the flow of genetic information in a cell is from DNA through RNA to proteins. This "Dogma" comes from the process known as transcription, a RNA copy of a segment of DNA, or messenger RNA (mRNA), is made. This strand of RNA can then be read by a ribosome to form a protein. Some RNA molecules are passive copies of DNA and often play crucial, active roles in the cell, e.g. switching genes on and off, critical protein synthesis machinery in ribosomes. [7], [12]

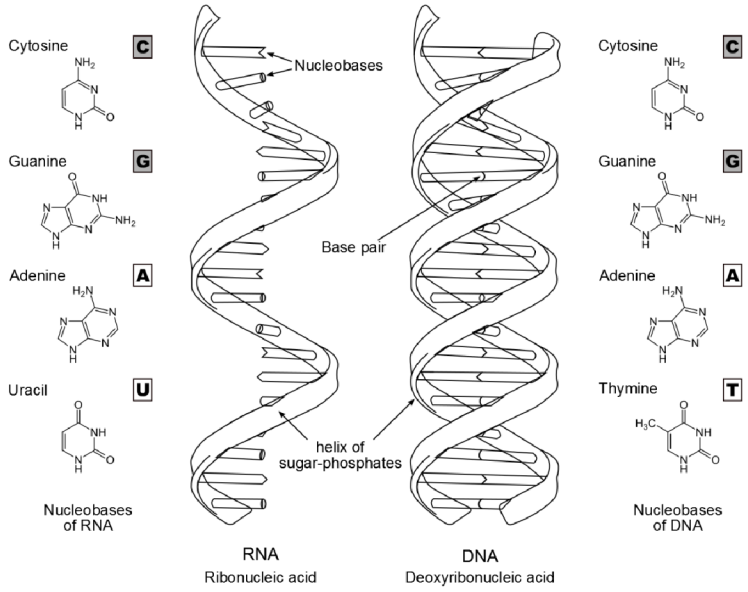


Figure 7 – RNA and DNA [48]

4.1. MicroRNAs

MicroRNAs (miRNAs) are the short endogenous non-coding RNAs, of about 22 (19-26) nucleotides in length, which modulate post-transcriptional gene regulation.
These molecules are involved in several cellular functions like apoptosis, cell development and differentiation, oncogenesis, tumour suppression and more others. The researches in this area established the involvement of miRNAs in various disease progressions, including certain types of cancer and researches shows that expression profiles of miRNAs are effective in classifying different types of human cancer. It is essential not only classify the diseases but also their molecular subtypes, because tumours might contain the discriminative signatures that are further propagated in the different forms of the disease.. It acts primarily by negatively regulating the expression of target mRNAs through translational inhibition and mRNA degradation. The complexity of post-transcriptional control of gene expression by miRNAs remains
a significant challenge. It has the potential to alter entire pathways due to their ability to target multiple genes simultaneously. miRNAs have been identified as prognostic various markers of breast cancer type and associated with breast tumours defined by their HER2 or ER/PR status. Approximately 50% of known human miRNAs are intronic, non-coding. miRNAs carry a unique signature that differs cancer subtypes and reveal new cancer subtypes. Generally, studies of miRNA are still big challenge and have a big importance, because understand can help in many fields of medicine. [24], [25]

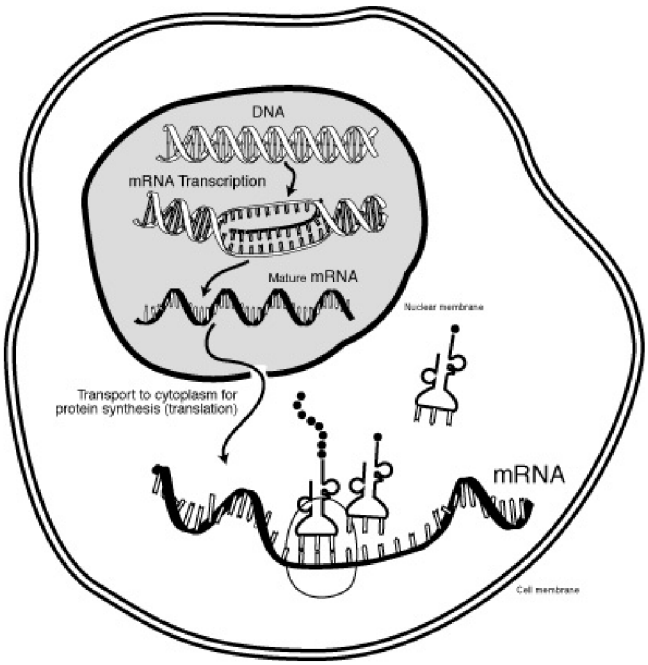


Figure 8 – Character and function of mRNA in cell [48]

4.2. Microarray

Modern biology and genomic sciences are rooted in infectious disease research. A comprehensive characterization of all of the genetic, genomic and epigenetic modifications associated with the cancer is critical for the understanding of the origins of tumour process and for finding the targets of therapeutic interventions. Microarrays is a technology for biology-molecular analysis designed to simultaneously monitor whole genome host and pathogen gene expression, providing a complete view of progression of an infectious disease state-how a pathogen responds to its host and the host to its pathogen. Utilization of high-density nucleic acid microarrays is one the most effective approaches to identifying these key molecular events. [19]

In the other words it is a laboratory tool used for measure the expression of large numbers of genes at the same time. Gene expression is the process when information from gene is used in the synthesis of functional gene product. For example for genetic code stored in DNA is interpreted by gene expression (transcription, RNA processing, translation).

The types of microarrays mainly depend on the company which produce the platform and also differ on the way of fabrication, accuracy, efficiency, cost, work. The most common used is microarrays from company Affymetrix, than Eppendorf and Illumina. In principle microarrays are microscope slides (glass or silicon) that are printed with thousands of microscopic spots in the defined positions with each spot containing a known DNA/RNA sequence or gene. The DNA molecules attached to each slide act as probes to detect gene expression or the set of messenger RNA (mRNA) transcripts expressed by a group of genes.

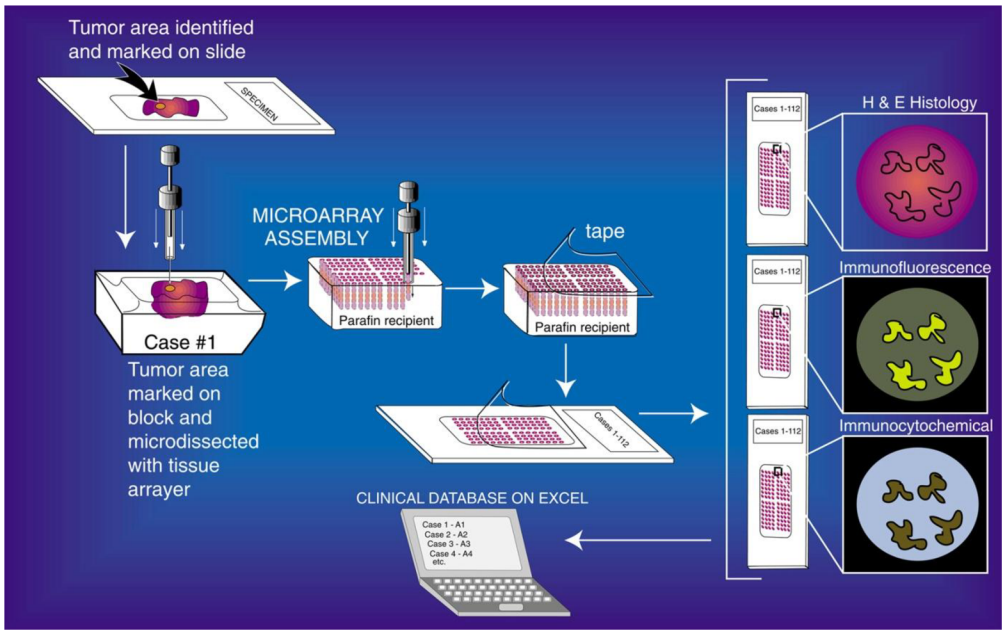


Figure 9 - Process of testing using microarray [48]

To perform a microarray analysis, mRNA test samples are typically collected from both experimental samples (from an individual with a disease) and a reference sample (healthy individual). They are then converted into complementary DNA (cDNA) and each is labelled with a fluorescent probe of a different colour (green, red). Then are both samples mixed together and allowed to bind the microarray slide. The process is called hybridization. Following hybridization, the microarray platform is scanned to measure the expression of each gene printed on the slide. Spot takes the colour of sample which expression is higher or it comes yellow if there is equal expression in the two samples. The steps of full process are shown in the figure 9.

Data gathered through microarrays can be used to create gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or treatment. These experiments are able to determine the relative expression of tens or thousands of genes in same time, this resulting with large databases and it is necessary to analyse this databases and take from it biologically relevant data, like the potential cancer biomarker genes, however it is not easy task across all different experiments, researches, microarray platforms or cancer types. Because of the large non constant variability of experiments and difference between microarrays techniques it is also important to find correct parameters of experiment to significantly affect the output of analyse which has biological and medical meaning. For the cancer research the analysis of microarray data the patterns of gene expression can be use for find a diagnosis or prognostic characterizing of a concrete illness stage or for detecting and proposing the role of specific genes in the cancer development i.e. detecting of cancer biomarkers. There are few methods for determination of potential cancer biomarkers, for example to cast the problem as multiple criteria optimization problem. [49]

Expect of different types of DNA microarrays we can also use other methods, like northern blotting, real-time polymerase chain reaction (RT-PCR) and serial analysis of gene expression (SAGE) or multiplex PCR. These tools, same like microarrays, allow the detection of differentially expressed genes, up- or down-regulated genes in relation to specific clinical conditions or functional pathways. Studies may be expanded or follow validation studies using additional gene expression data measured with alternative experimental platforms or other “omic” approaches. [15], [16]

An important goal in cancer research is to identify significant genomic alterations responsible for the emergence and progression of disease. With Microarray it is now possible to perform extensive analysis of tumour genomes.

In other words the goal of cancer research is ultimately to improve the diagnosis and the treatment of cancer through more accurate disease classification and patient stratification, which allows for the design of therapies that are more targeted to specific

cancer subtypes and potentially improves effectiveness of existing regimens based on therapeutic response and adverse events. [18]

One third of human miRNA host genes are hybritized by probes on U133plus2 Affymetrix gene chip. Many intronic miRNAs show significantly correlated expression profiles with their host genes. 70% of miRNAs has expression profiles are significantly correlated with their host gene. The expression of these miRNAs can be inferred from the expression of their host genes and can be evaluated as putative prognostic markers in breast cancer and its subtypes using gene expression data.

Problem is that the number of features in the microarray datasets can be tens of thousands but the number of cases rarely exceeds a few hundred and often is less than one hundred. So it is for what were involved methods which are available to work with p>>n class prediction problems. [11]

4.3. Transcriptomic datasets

Breast cancer is a complex heterogeneous disease for which a substantial resource of trancriptomic data is available. It has traditionally been sub-classified depending, amongst other factors, on the expression of different receptor proteins, such as estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). These “biomarkers” allow us to tailor the level of clinical interventional. While ER-positive the second positive should be deleted tumours receive hormone therapies and HER2-positive cancers receive targeted therapies, “triple negative” cancers lacking these markers currently have no targeted therapies and cause a disproportionate number of breast cancer deaths.

In addition to the traditional classifications using these biomarkers, in recent years, whole genome DNA microarrays have been utilised to further classify this disease, initially into five molecular subtypes based on gene expression profiles Luminal A and luminal B (ER-positive tumours), HER2 (HER2-positive tumours), normal-like and basal. It is important to identify which breast cancer patients are at risk of developing a more aggressive phenotype so as to tailor the level of clinical intervention. Prognostic biomarkers can be used to assess the inherent likelihood of a patient exhibiting a particular outcome. There is a wide spectrum of survival requiring the identification of additional novel prognostic markers. Triple negative subtype has no such prognostic biomarkers currently in clinical use. But still there is a great deal of transcriptomics data currently available to facilitate the identification of novel molecular biomarkers associated with breast cancer and its subtypes. [11]

5 Optimization methods

In computer sciences, dynamic programming is an approach to solve faster certain kinds of problems because they can be split into sub-problems which overlap. Main idea of the optimization methods is to solve complex search problems with discrete optimization concepts and algorithms, including constraint programming, local search and mixed interreges programming. Optimization methods are designed to overcome inverse problems. The first roots we can find in 19th century in economy work of Edge worth and Pareto.

Optimization problems are seldom single-objective. Usually, there are several contradictory criteria or objectives that must be satisfied simultaneously. Multi objective optimization (MOP) is a discipline centred in the resolution of this kind of problems. The techniques can be classified into exact and approximate (stochastic and heuristic) algorithms. MOP (single or multi objective) can be divided into two categories. Those whose solutions are encoded with real-valued variables (known as continuous optimization problems) and solutions which are encode by using discrete variables. These problems are usually called Multi objective combinatorial optimization problems (MCOP). Most of Metaheuristics for solving MOPs are designed to deal with continuous type of problems. However many real problems are MCOPs. [1], [3]

5.1. Discrete optimization methods

Discrete optimization forms a class of computationally expensive problems of significant theoretical and practical interest. Algorithms systematically search the space of possible solutions. [1]

A discrete optimization problem (DOP) can be expressed as a set of data and function (*S*,*f*). The set *S* is a finite or countable infinite set of all solutions that satisfy specified constraints. The function f is the cost function that maps each element of the set *S* into the set of real numbers *R*. The objective of a DOP is to find a optimal solution *xopt*, such that $f(xopt) <= f(x)$ for all *x* belongs *S*. The feasible space *S* is typically very large. For this reason, a DOP can be reformulated as the problem of finding a minimum-cost path in a graph from a designated initial node to one of several possible goal nodes. Each element *x* in *S* can be viewed as a path from the initial node to one of the goal nodes. This graph is called a state space. Often it is possible to estimate, so-called heuristic estimate, the cost to reach the goal state from an intermediate state, it can be effective in guiding search to the solution. If the estimate is guaranteed to be an underestimate, the heuristic is called an admissible heuristic. Admissible heuristics have desirable properties in terms of optimality of solution. [1]

Application is in a number of diverse problems such as VLSI layouts, robot motion planning, test pattern generation and facility location can be formulated as DOPs.

Also other more common problems or interest like a roll cutting at paper mills, 2D board cutting, terminal location, production planning, routing/scheduling, game (Sudoku) we can subsume between problems which can be resolved with help of Optimization methods. [2]

5.2. Greedy Optimization

Greedy algorithm is an algorithm that follows the problem for solving heuristic problems in the way to make the locally optimal choice at each stage with the hope of finding a global optimum.

In many problems, a greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time.

To construct the solution in an optimal way algorithm maintains two sets. One contains chosen items and the other contains rejected items.

The greedy algorithm consists of four functions.

1. A function that checks whether chosen set of items provide a solution.
2. A function that checks the feasibility of a set.
3. The selection function tells which of the candidates is the most promising.
4. An objective function, which does not appear explicitly, gives the value of a solution.

Unlike Dynamic Programming, which solves the sub-problems bottom-up, a greedy strategy usually progresses in a top-down fashion, making one greedy choice after another, reducing each problem to a smaller one. The limit approaches in selecting molecular signatures.

Dynamic programming is effective for problems of small size (same like methods such as brand and bound).

Multi objective functions whose optima can be found are very rare. The reason why the optimization problem could be solved is that the two objectives of the function were separable. Greedy is a strategy that works well on optimization problems with the following characteristics which are Greedy-choice property (A global optimum can be arrived at by selecting a local optimum) and Optimal substructure (An optimal solution to the problem contains an optimal solution to sub-problems). [3], [4], [5]

6 Meta-heuristic methods

Meta-heuristic is a rather unfortunate term often used to describe a major subfield of stochastic optimization. Stochastic optimization is the general class of algorithms and techniques which employ some degree of randomness to find optimal solution to hard problems. It is applicable to a very wide range of problems. Meta-heuristic are used for a harder problems, approximate algorithms are mandatory. It makes far loose assumption, against of classical Optimization methods, and sometimes makes none at all.

Meta-heuristics share a number of common characteristics concerning in particular following two elements encoding and fitness evaluation. Encoding is the way how to represent the candidate solutions of the search space and fitness function is the way of measuring the quality of the candidate solutions. This evaluation function introduces an order among the solutions of the search space, allowing thus the comparison of pair solutions. It provides just a general optimization framework that can potentially be applied to various search problems. It must be carefully adapted to the given problem and integrate problem-specific knowledge. Techniques which constitute meta-heuristics algorithms range from simple local search procedure to complex learning processes.

Not a commonly accepted definition of meta-heuristics is that they can be considered as high-level strategies that guide a set of simpler heuristic techniques in the search of the optimum. For the problems with more than two criteria, there are no many effective exact procedures, due the simultaneous difficulties of NP-difficult complexity and the multi-criterion nature of the problems.

Among these techniques, evolutionary algorithms for solving MOPs are very popular, giving rise to a wide variety of algorithms. For example hill-climbing method is a simple meta-heuristic algorithm. It starts with random behaviour set. Then it makes a small, random modification to it and try the new version, if it is better than throw the old one away. If the newest version is better, throw away the current version and also throw away the newest version. It is repeated as long as it is possible. The algorithm can be also little bit more aggressive with utilise of some modifications. Hill-climbing exploits a heuristic belief about your space of candidate solutions which is usually true for many problems. That similar solutions tend to behave similarly, so small modifications will generally result in small, well-behaved changes in quality, allowing us to “climb the hill” of quality up to good solutions. This belief is one of the central defining features of meta-heuristic. [3], [15]

6.1. Division of Meta-heuristic methods

The solving methods can be divided into two classes, algorithms which are specific to a given problem, neighbourhood-based local search methods and Meta-heuristics, which are applicable to a large variety of MCOPs.

There is more ways of dividing Meta-heuristics depends on literature and way of look on it. But generally one type of strategy is an improvement on simple local algorithms, this include simulated annealing, Tabu search, Scatter search. Other type has the learning components to research. Other can be the classification dimension a single solution or population/based research. [22]

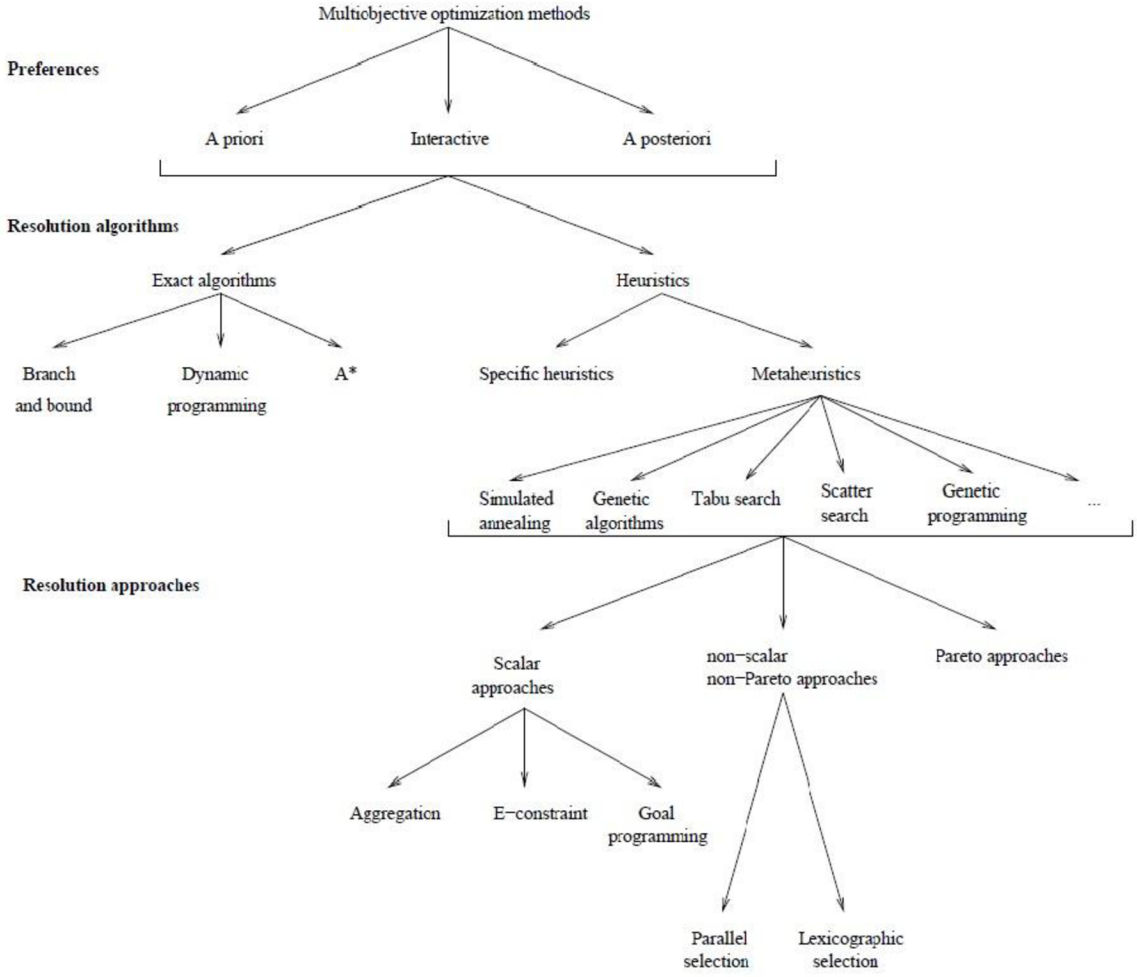


Figure 10 – One of the examples of dividing meta-heuristics optimization problems

The approaches used for MCOPs resolution can be classified in three main categories.

First is Scalar approaches. This method implies the transformation of the MCOP from multi-objective problem into a single-objective one. Algorithms based on aggregation, which combine the various cost functions into only one objective function

F. Require for the decision maker to have a good knowledge of its problem. The Scalar approaches include Aggregation, E-constraint and goal programming.

Second category is a Pareto Approaches. It is based on use of the concept of Pareto optimality in their search. The process of selection of the generated solutions is based on the concept of non-dominance. And last one Non-Pareto and non-Scalar approaches, which are the operators to treat the various objectives separately. These approaches mostly based on populations of solutions, the research is carried out by treating the various non-commensurable objectives separately. We can fractionate them to two selections use for Genetic algorithms. It is Parallel selection and lexicographic selection. We can also create hybrids of various meta-heuristic algorithms. [1], [15]

The Genetic algorithms are the one of the most growing and in cynosure they are the methods of Meta-heuristic. The point of population-based methods is to inspirit themselves with the concepts from biology. One set of techniques, known as Evolutionary Computation borrows liberally from biology, genetics and evolution. It generally resembles techniques. An algorithm chosen from this collection is Evolutionary Algorithm (EA). Most EAs may be divided into generational algorithms, which update the entire sample once per iteration, and steady-state algorithms, which update the sample a few candidate solutions at the time. EAs include the Genetic Algorithm and Evolution strategies. Each of them have generation and steady-state version. New samples (populations) are generated or revised based on the results from older one. [3]

In addition to the algorithms above, there are hybrid and parallel meta-heuristics. Parallel meta-heuristic use the techniques of parallel programming to run multiple searches in parallel.

A hybrid meta-heuristics combines a meta-heuristic with other optimization approaches, as algorithms from mathematical programming, constraint programming and machine learning. Both components of this method are running concurrently to exchange information to guide the search. [22]

7 Machine Learning

Machine learning (ML), as a sphere of artificial intelligence, is the discipline to concern the design and development of algorithms that allow to systems to evolve behaviours based on empirical data, such as sensor data or databases, in the case of bioinformatics.

Main task of the ML is to extract useful features from given data and build a statistical model over this data. It arises from the need of automated or artificial systems to have possibility to find a decision based on a given model. [20]

The objective of a system of pattern recognition is either to estimate a real value (regression) or to estimate an appropriate label (classification) corresponding to a given input data. The model, the name of this method to cue, is based on some knowledge acquired by training on the select learning data. This method has huge use in the application on the problems from the real world. The spectrum of application is very broad. ML is useful in fields like data mining, text categorisation, biomedical problems as data analysis, Magnetic Resonance Imaging, signal processing, automatic speech recognition, speaker identification, character recognition, diagnostic and system monitoring and decision, page ranking and image processing. And more recently and in last year's increasing number of application is in the field of biomedical engineering. For example we can to mention Biocomputing or DNA sequence identification, Automatic analysis of digital mammography or electrographs and biometrics, which contain from personal identification based on biological data or neurosciences, which are also strong branch of ML methods. And there is growing number of applications in home-machine interface and behaviour analysis, like Driver behaviour analysis, elderly and disabled behaviour analysis at home or brain/computer interface. [20], [21]

There are two types of ML classification algorithms supervised and unsupervised learning; their schemas are in figure 11.

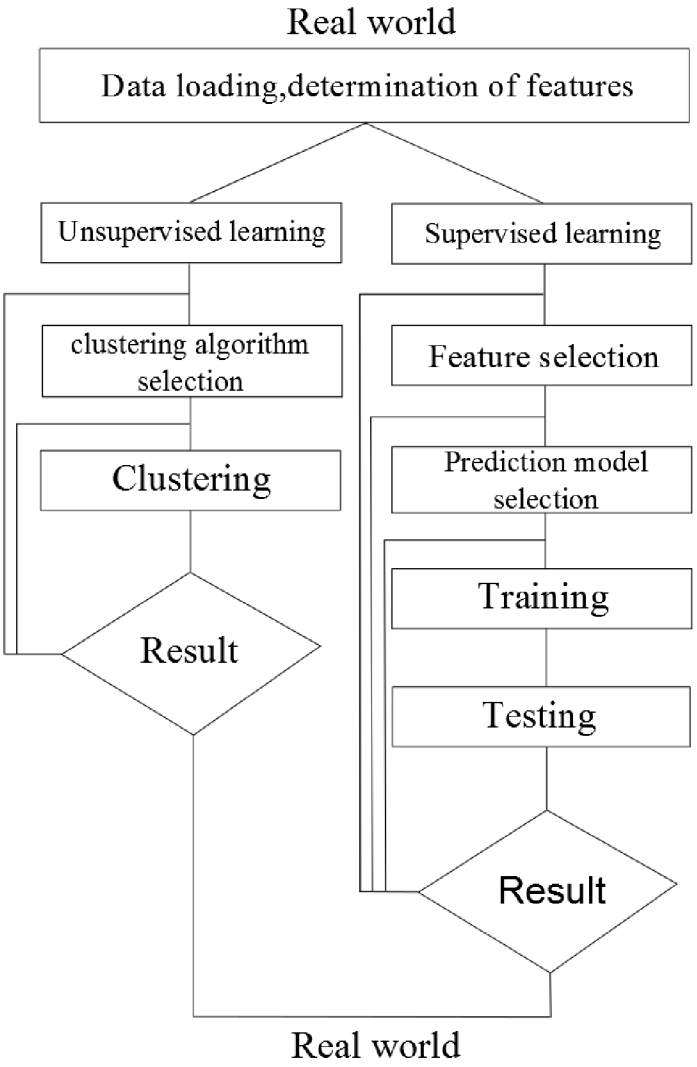


Figure 11 - Supervised vs. unsupervised classification

Supervised algorithms use a set of available labelled training data, so the class labels are known beforehand and it is also known the correct output should be, so there is the relationship between input and output. Supervised problems are categorized into "regression" and "classification" problems. In the regression problem the prediction of the results is within a continuous output, mean that input variables are mapped to some continuous function. In the classification problem on the other side is used discrete output for prediction of results, so the input variables are mapped into discrete categories. Unsupervised learning is use a set of available unlabeled or partially labelled training data, so there is no idea how the results should look like. The structure is derived from data because there is no necessity to know the effect of variables, so there is no teacher to correct the results, i.e. no feedback. To this category belong for example associative memory or clustering methods, where the structure is derived based on relationships among the variables in the data. [21], [39]

In general, the main problem of classification is to find the right function that can maps an Rx1 input feature vector to a class label in which the information about data is encoded in an appropriate way. When the classification problem is defined, there is the big variety of mathematical tools, like optimization algorithms to be used to build a model. [21]

In general the machine learning methods consist from few steps to receive ideal prediction. Three frequently occurring kinds of objectives which work with thousand of features for each case in DNA microarray investigation are class comparison, class prediction and class discovery. [20]

Class comparison means identification of differentially expressed genes in cells from different type of tissue of different kinds of patients or in cells exposed to different experimental conditions. The characteristic feature is that the classes which are defined during the process should to be defined independently of the expression data. It can identify the genes that are differentially expressed between patients who respond to a specified treatment and those who do not respond.

In class prediction are also classes defined independently of the expression data. The point of the class prediction is in developing a multi-gene classifier that can be than applied to expression profiles of samples whose class is unknown. The main aim of class prediction is to develop a classification function it can be used to predict if every new patient will respond to the therapy based on the gene expression profile of the tumour. It has use in the medical problems of therapy selection, diagnostic classification or prognostic prediction.

In class discovery there is no classification defined independently of the expression profile. In this method the objective is to discover subsets (clusters) of the cases revealed by gene expression profiles and to identify the genes that distinguish the clusters or to discover classes of co-regulated genes. [29], [31]

8 Class prediction

The four main components to developing a class predictor are Feature selection, selecting a prediction model, fitting a prediction model to training data and estimating the prediction error that can be expect in future use of the model with independent data.

8.1. Feature selection

Feature selection (FS) is trying to give the answer for the question which features should be used for create a predictive model, which are provide the most discriminatory information. It is the key to developing an accurate class predictor. FS can be applied to both supervised and unsupervised learning. [23]

The aim of feature selection techniques are to select the minimum set of relevant highly informative attributes for gain the relevant results, i.e. method is looking for a subset of features that leads to the best generalisation performance of the classifier when trained on this subset. The function is than often the efficiency of a subsequent classifier trained on the given set of this features. It is the apparent need in lot of biomedical applications, some literature subsume FS between techniques of sequence, microarray and spectral analysis. [20], [23]

The subset of features can be different for every method. The solution of this problem can be to evaluate for every subset of features by training a classifier with each subset, observing performance and select the subset with the best performance.

In contrast to the other dimensionality reduction techniques, like principal component analysis or information theory using, FS do not change the original representation of the variables, only select a subset of them. This is the main advantage of these methods because it preserves the original significance of the variables. But it means also disadvantage that we need to find the optimal model parameters for the optimal feature subset, even there is no guarantee that the parameters are optimal for full feature set as for the optimal feature subset. [20], [23]

The next paragraph is focused on supervised learning; the FS for unsupervised learning is more complex.

The intentions of FS are various. The most important are to avoid the over fitting and improve performance of the model, to provide faster and more effective models and to gain deeper into the underlying processes for data generation. The FS techniques differ from each other in the way of finding the optimal subset of relevant features in the model hypothesis space in the model selection. This subset, named feature vector (p) form the classifier and is the collection of r scalar or matrix of the representative classes in the feature space R, where $r<R$. [21], [22], [23], [31]

The methods for FS can be divided to three groups, accordingly how they combine the feature selection with the construction of the model classifier.
The categories are filter methods, wrapper methods and embedded methods.

The characteristics of the individual advantages and disadvantages and examples of using techniques are shown in table Figure 12.

The filter techniques assess the relevance of features by looking at the intrinsic properties of the data. The score is calculated and low-scoring features are removed and the rest of features are the subset which is presented to the classification algorithm, so the feature subset is independent on the model selection step.

On the other side wrapped methods embed the model hypothesis search within the feature subset search. The various subsets of features are generated and evaluated in the defined space of the possible variables. The evaluation of the specific subset of features is acquired by training and testing a specific classification model.

The algorithm is wrapped around the classification model to find the space of all feature subsets. Wrapper-based feature selection algorithms can be used only if the features are uncorrelated and independent on each other and for large number of features. [29]

For example hill-climbing method, simulated annealing, genetic algorithm, greedy forward selection, greedy backward selection, or more other efficient methods (depth-first search, branch and bound search and others) of search algorithms competence to this group can be used, but each has its own limitations. Heuristic search methods are used to search the optimal subset of features, they depends on the size of the space of features subsets, which grows exponentially with the number of the number of features. These methods can be divided into two classes Deterministic and Randomized search algorithms.

The last category is embedded techniques, where the search for the optimal subset of features is build into the classifier construction, i.e. combine space of features subset and hypotheses. These methods are also specific to a given learning algorithm.

Advantages and disadvantages of all three classes are shown in Figure 12. [21]

Model search	Advantages	Disadvantages	Examples
Filter	**Univariate**		
	Fast	Ignores feature dependencies	Euclidean distance
	Scalable	Ignores interaction with the classifier	i-test
	Independent of the classifier		
	Multivariate		
	Models feature dependencies	Slower than univariate techniques	Correlation-base feature selection
	Independent of classifier	Less scalable than unovariate techniques	Markov blanket filter
	Better computional complexity than wrapper methods	Ignores interaction with the classifier	
Wrapper	**Deterministic**		
	Simple	Risk of over fitting	Beam search
	Interacts with the classifier	More prone than randomized	
	Models feature dependencies	Classifier dependent selection	
	Less computationallz intensive than randomized methods		
	Randomized		
	Less prone to local optima	Computationally intensit	Simulated annealing
	Interacts with the classifier	Classifier dependent selection	Genetic algorithms
	Models feature dependencies	Higher risk of over fitting than deterministic	
Embedded	Interacts with the classifier	Classifier dependent selection	Decision trees
	Better computational complexity than wrapper methods		Feature selection using the weight vector
	Models feature dependencies		

Figure 12 - A taxonomy of feature selection techniques

8.2. Application of feature selection methods

Next to other applications for this work is important application for Bioinformatics for microarray analysis.

The goal of microarray data, which is giving a great challenge for computational techniques, is their large dimensionality, i.e. several tens of thousands of genes and small sample size. So there can be more noisy variables than relevant one, i.e. the noisy variable that is not related to the thing being predicted. And thanks the high amount of the noisy variables the prediction can be less accurate, how is known from the theory of linear regression. To use right feature selection can help not to lose the genes with the good influence for class prediction.

The univariate filter techniques dominate because their outputs are intuitive and easy to understand and could fulfil the objectives and expectations to validate the results coming from the laboratories or literature searches, also good for select the genes in a multivariate way for other techniques like data analysis techniques and less of time needed.

For identifying of different expressed genes can be used simple heuristics. They can be divided to three groups parametric, non-parametric and model-free methods. Parametric methods assume the given distribution from which the samples are generated. The most used for microarray analysis are t-test and his modification and ANOVA. Modifications of t-test better deal with small sample size and statistics. Model-free methods are frequently borrowed from statistics as Wilcoxon rank-sum test and some other specific methods. For multivariate gene selection are mainly used combinations of wrapped and embedded methods. [23], [29]

The most commonly used approach to FS is to identify differentially expressed genes among the classes when considered individually. For two classes, for example compute t-test or Mann-Whitney Wilcoxon test for each gene. For the entry to the class predictor are than chose the genes which are differentially expressed at the specified significant level. There is lot of the methods from several authors which provides the good discrimination of the classes.

We can use the technique of cross validation to provide the performance of the true system performance on the new data with using test data. [21], [29], [31]

8.3. Modelling and fitting prediction model to training data

The main aim of this step is to choose the type of classifier and a suitable training algorithm. Training is the procedure of classification when the classifier learns relationships between feature vectors and their labels or regression values, so training algorithm is built using set of labelled feature training data. On the other side we need to have separate validation and test data for evaluation. This are generated by the learning algorithm and are used, which are generally collected in the same time or took from

existing labelled data. Finally we can apply our classifier on unknown or unlabeled data to classify their class.

Generally classifier (model) is the system with parameters to find the right decision borders over training algorithm. When data are labelled incorrectly we are talking about cost of error, which measure the cost of taking decision. There are two types of cost of error, false positive and false negative. False positive case means that for example patient is healthy but the classifier predicts that he or she is sick. False negative case on the other side not to recognize the warning signs of sickness and declaring that the patient is perfectly health, even it is not right decision. The cost of this type of error could include death.

There are two other parameters for evaluating the performance of the training system like training performance and generalisation. Training performance respond to the performance of the classifier to be able correctly identify the classes of the training data during the training phase, this performance is not a good indicator of the most significant performance. Generalisation on the other side shows performance of the classifier to identify the classes of new data that were not introduced in the training phase. [21], [22], [29]

There is a lot of classifiers for training-based classification, here are mention just few of the most common ones. The first group are statistic-based classifiers (Statistical pattern recognition), this type of classifiers methods include Bayes classifier, Naive Bayes Classifier, Linear and Quadratic Discriminant Analysis, Support vector machine, hidden Markov models, nearest neighbour. Other groups are artificial neural networks and Decision trees.

Obviously there is a lot of different classification algorithms, every with other advantages, disadvantages, specification and features. Because during the research is used LDA, next chapter is going to put more light especially on this concrete method. [31]

8.4. Linear Discriminant Analysis

LDA was at first developed for two class problems but later it was transform ate also for problems of several classes (multivariate cases).

A major role of linear classification or say multivariate analysis is to find a transformation of multivariate data, to find ideal classifier fitted and trained on the given set of features that reduces the data set and make it easier, smaller and better representative to understand or analyse. So many significant variables are hidden in the original data. Because of this the main deal of LDA is to find a subset of features that leads to the best generalisation performance of the classifier which is trained on thus training subset. This best subset becomes a function of the classifier. However, the new variables can represent better the process under original data.

The predictor is fitted to a set of data; parameters of predictor must be specified proportionally to the number of genes selected for inclusion in model.
The objective is to find a transformation such that between-classes distance are maximised and the within-classes distances are minimised in the transformed space. These distances are measured by using the between-class scatter matrix and within-class scatter matrix. In other words the main idea is to find a line, hyperplan or hypercube projections such as the samples of different classes are well separated.
The original distribution of classes should be unimodal, in other case LDA becomes ineffective. [21], [22], [29]
A linear discriminant is a function is f(x) (1)

$$f(x) = \sum_{i \in F} w_i x_i \tag{1}$$

where x_i denotes the log-ratio or log-signal the i'th gene,
w_i is the weight given to the gene and the summation is over the set of features (genes) F selected as the input for the class predictor.
In the case of two classes there is a threshold value d and the prediction is to class 1 or class 2 depending if computation of f(x)(1) is lower or greater than d respectively. [29]

In real, the means and covariances of a given class are unknown. But there is possibility to estimate these values on the basis of training. The problem in LDA is coming when the number of features is greater than the number of instances in the class. Than the estimation of the covariance does not have full rank and cannot be inversed.
Other thing is that not even ever covariance can be considered as optimal and so results. The measuring of the separation between the projections of two classes is over their mean scalars μ. There is assumption that the greater the difference in $|\mu_1-\mu_2|$ is the better is the separation between classes. In the process of measurement it should be also account with the variance of classes and with need of normalization $|\mu_1-\mu_2|$ by a factor which is proportional to the variance. Also it is important to define the scatter matrices S_w, proportional to the covariance matrix of the original data of dimension R_0, to define the spread in the original data. When there is one of the previous problems the one of the method to solve it is to use a pseudo inverse instead of the usual inverse matrix S_w or use the Shrinkage estimator. For two classes it have to be defined and measured the between class scatter matrix (2)

$$SB = (\mu 1 - \mu 2)(\mu 1 - \mu 2)^T \tag{2}$$

which measure the average spread of the 2 classes before the projection. Measure the objective function. Similar process is using for multivariate classes cases. [20], [21]

8.5. Estimating prediction accuracy

Performance evaluation coming once the model selection and training is completed. Its generalization performance needs to be evaluated on previously unseen data to estimate its true performance on field data. One of the methods, the most popular one, is to split the data into two parts, when the first one is used as representative training set R and second is used for testing set N the performance of the algorithm. Important is to find ideal size of both sets, there should be almost equal, often N < R.

Knowing that there are statistically significant genes is not enough. It is important to know how accurately is possible to predict in which class the new sample to come under.

For this information is found the fully specified predictor help to the put aside set of test samples, i.e. to properly estimate accuracy of a predictor the working set of samples must be partitioned into a training and test set.

The test set represent the set of the samples for which class labels are to be predicted, so it cannot be used for the development of the prediction model and for the selecting of genes to be used in the model. One of this methods is split-sample method of partitioning the set of samples into a training and test set easily, for example two third of samples as training set and the rest as test set. Other alternative to the split sample method is the cross validation (CV). There are more types of CV k-fold cross validation, leave-one-out validation. In k-fold CV are data split, in *k* iterations, into k>2 partitions to create *k* blocks of data. Of these blocks, *k-1* is than training data and remaining *kth* block is used for testing. During every iteration the procedure use a different block for learning and testing. The advantages of this method are that all data are used for training same for testing, but never in the same time and because method is repeated *k* times, the probability of mistake is reduced through averaging. In leave-one-out validation, the entire data but one is used for each testing session and then testing on the remaining cases, repeating N times. Problem of this method is the worst confidence interval. [21], [29], [31]

9 Statistical methods

Statistical methods such as supervised classification and machine learning identify distinguishing features associated with disease subtype but are not necessarily clear or interpretable on a biological level.

9.1. Statistical validation of prediction

If clinical predictions are synthesized statistically it is statistical prediction. "Clinician" can integrate the output of a statistical prediction scheme and his or her own judgments. It can be synthesized that two or more potentially disparate items of information are available. Then clinicians must use their own judgment to deal with discrepancies in the data. The statistical formula requires no professional judgment to arrive at a prediction, statistical data combination. Normally results from experiments are compared with public datasets in oncology. It is about comparison of two classes of samples to detect statistically important differences between gene expression and which aimed to support the prediction of disease emergence or progression. [16], [13]

There exist huge amount of statistical methods. In variable selection the standard approaches to use are two-sample t-test and Wilcoxon-Mann-Whitney test (WMW) because the aim is to know which of two groups has generally larger responses. Both of them are usually associated with clearly defined set of different hypotheses, the decision rule and p-value can be associated with other sets of assumptions. These perspectives associated with p-value allow different interpretation of this p-value.

In contrast to t-test, the Wilcoxon test is robust against outliers, which are frequent in microarray and does not require normal distribution of the expression levels within both classes which is often questionable, it is why in the research is used WMW test, next chapter introduce this statistical method more deeply. [34], [38]

9.2. Wilcoxon-Mann-Whitney test

In the statistics WMW test is the alternative test to the independent t-test. This test is non-parametric and compares two population means which come from the same population. Also can be use for test if two population means are equal or not. The size of samples for the test has to be same. The Wilcoxon-Mann-Whitney has few alternatives like Mann-Whitney U test for ordinal data, Wilcoxon rank sum test and Kendall's test, which is similar to Mann-Whitney U test and can be equivalent to the chi-square test.

The sample for the test drawn from the population is random and independent within the other samples and do not affect each other. Using WMW we can decide if the population distributions are identical without assuming them to follow the normal distribution.

The equation for measurement is in (3)

$$U = n_{1n_2} + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i \quad (3)$$

Where n_1 is the sample size one and n_2 sample size two, R_i is than the rank of the sample size. [36], [37], [38]

9.3. P-value

P-value is obtained from the various statistical tests, let say it is statistical comparison procedure which can say a lot about statistical data. It helps to determine the significance of results during performing a hypothesis test in statistics. It is the way of counting probability associated with finding differences in the populations being compared.

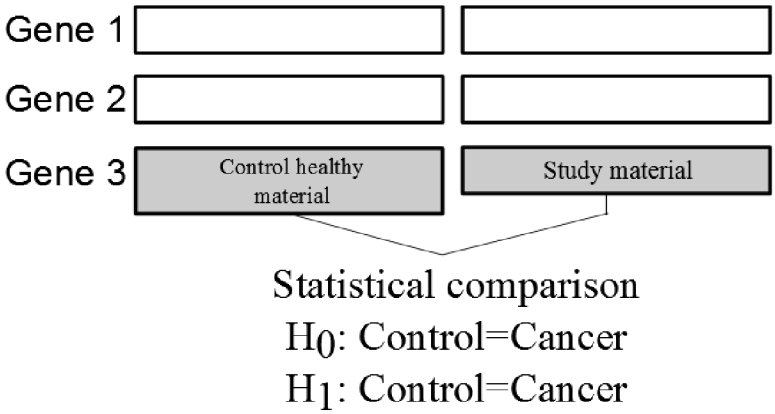


Figure 13 - Schematic example of how to obtained p-value [49]

Figure 13 shows the way how to obtain the p-value from expression level for every gene in given tissue.
The criterion is performance is

$$P(i) \leq \frac{\alpha}{q-i+1} \quad (4)$$

Where q is number of total hypothesis and α wise error rate, P(i) correspond to the number of gene under evaluation. For α usually choose value $\alpha < 0.05$.
Hypothesis tests are used to test the validity of a claim that is made about a population, this claim is called the null hypothesis. The alternative hypothesis is the one we are believe if the null hypothesis is concluded to be untrue. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence.

P-value is a number between 0 and 1 and is interpreted in the following way:

- The lower p-value (typically <= 0.05) indicates strong evidence against the null hypothesis, it means the large difference between comparing groups, so stronger evidence of statistical significance. It is the more significant to have a change in level of expression.
- A large p-value (> 0.05) shows opposite, so it fail to reject the null hypothesis
- P-values very close to the cut off (=0.05) are considered to be marginal.

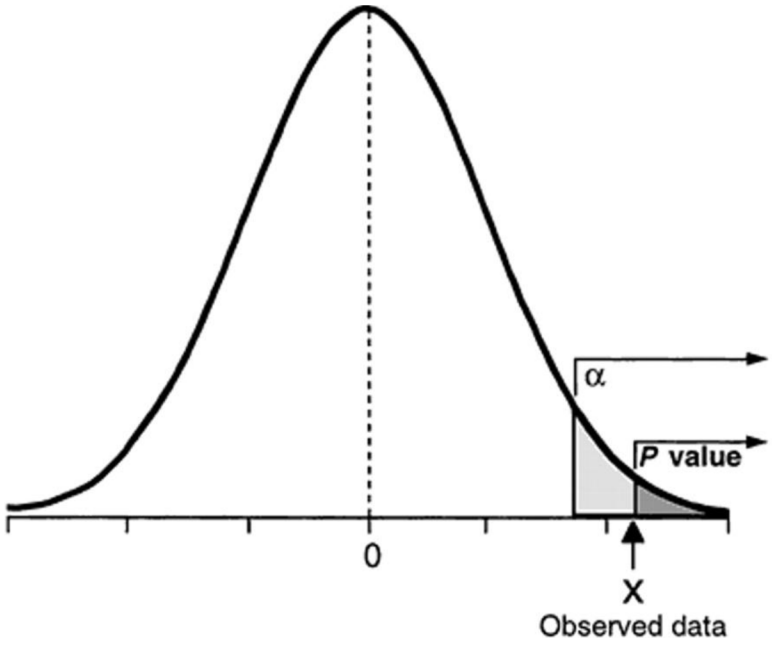


Figure 14 - P-value distribution

Each gene with a value of performance measure is going to be minimized. The smaller the p-value is the more important is the change of expression of the concrete gene under consideration, so for to have a good results small values are the most attractive.
The p-value counting is also useful for do statistical comparison procedure to get the two parameters from both two states of tissues cancer one and healthy one for a particular gene. [49]

9.4. Standard deviation

SD can be difficult to interpret as a single number on its own.
A small standard deviation means that the values in a statistical data set are close to the mean of the data set, on average, and a large standard deviation mean that the values in the data set are farther away from the mean, on average. Generally the standard measures how concentrated the data are around the mean; the more concentrated, the smaller the standard deviation.

A small standard deviation can be a goal in certain situations where the results are restricted. A large standard deviation is not necessarily a bad thing. It just reflects a large amount of variation in the group that is being studied.

9.5. Accuracy and robustness of statistics

To compare classification performance there can be use computing of sensitivity (true positive rate), specificity (true negative rate) and accuracy (AC). Sensitivity (SE) measure the proportion of actual positives which are correctly identified as such complementary to the false negative rate. Specificity (SP) is on the other side true negative rate which is noticed as the false positive rate. The ideal predictor has 100% sensitivity and specificity.

Sensitivity is compute as

$$SE = TP/(TP + FN) \quad (5)$$

Specificity as

$$SP = TN/(TN + FP) \quad (6)$$

Accuracy as

$$AC = (TP + TN)/(TP + FN + TN + FP) \quad (7)$$

Where TP is number of true positive (Samples with disease correctly diagnosed as sick), TN is number of true negatives (Health samples correctly identified as healthy), FP in number of false positive (Samples without disease incorrectly diagnosed as sick) and FN in number of false negatives (Sick samples incorrectly diagnosed as healthy). [33]

10 Bimodality index

The Bimodality index belongs to statistical methods. Identifying genes with bimodality index from large-scale expression data is important analytical task. Bimodality index, not only identify but also rank meaningful and reliable patterns. BI can be compute with use of the techniques as K-means, model-based clustering, tests of bimodality, mixture model-based algorithm, Dip test, Kurtosis, Markov chain Monte Carlo technique or Bimodality Index algorithm (BI). In oncology this way maybe can be part of search for clinically important therapeutic targets inside tumour and for classification of tumour subtypes to help understand clinical and biological character of cancer.

BI works with different expression, with two modes centred on the mean expression of a gene in two distinct subgroups of samples. The input is than the mixture of two populations with distinct means and the samples are characterized belong to each of two distributions. For example two components normal mixture-model-based clustering algorithm using clustering methods and represent each component as a cluster. This algorithm provides also more statistical data as mean, standard deviation or sample proportion. Problem is to find appropriate test standard to estimate p-values and chose suitable cutoff for this can be adopt likelihood ratio test with hypotheses of bimodal or unimodal distribution. It is also often combine with Akaike information criterion or Bayesian information criterion or with identification of genes to the assumed mixture distribution. K-means method use Clustering algorithms, same like model-based clustering. Dip test is defined as the maximum difference between an empirical distribution function and the unimodal distribution function that minimizes that maximum difference. Other method for find bimodality expression is method (Andrew et.al) Profile Analysis using Clustering and Kurtosis (PACK) which is using
the expectation-maximization algorithm and Kurtosis to tag major and outlier bimodality pattern. This method is successful in use for finding biomarkers.

The algorithm of Bimodality index, found by Wang et al. takes a criterion to identify and rank bimodal signatures from gene expression data that provides bimodality for fine distinctions between the bimodality expressed genes and give
a continuous value of bimodality for each gene not only information if bimodality exist or not. It is possible than rank genes and focus the interest on those with the strongest evidence of useful and more reliable bimodality patterns.

For a gene with bimodal expression, the distribution can be expressed as
a mixture of two normal distributions

$$y = \pi N(\mu_1, \sigma) + (1 - \pi) N(\mu_2, \sigma) \tag{8}$$

Where π is a proportion of samples in one group,
μ_1 and μ_2 are the means of the expression level of two modes and σ is common standard deviation. y is than the expression measurement.

$$BI = [\pi(1-\pi)]_2^1 \delta \quad (9)$$

The equation BI in (9) computes the bimodality.
Where $\delta = |\mu_1 - \mu_2|/\sigma$ is the effect size that measures the distance between the two-components. π and δ can be estimated for a given dataset. A limitation of this method is that it is defined based on a normal mixture with equal variance not for other possibilities.
Values BI are the distributions that are equally separable. Larger BI than shows smaller sample sizes and bimodality representation is easily distinguishes.
Compare to other methods may these described in previous article or others it seems that Wank's bimodality index algorithm has really good results in the introduction of finding meaningful and reliable bimodal indexes. Wang than using cut-off BI=1.1 to show on the genes with higher bimodality, which should be n his opinion bimodal, so useful biological markers. [40], [41], [42], [43], [44]

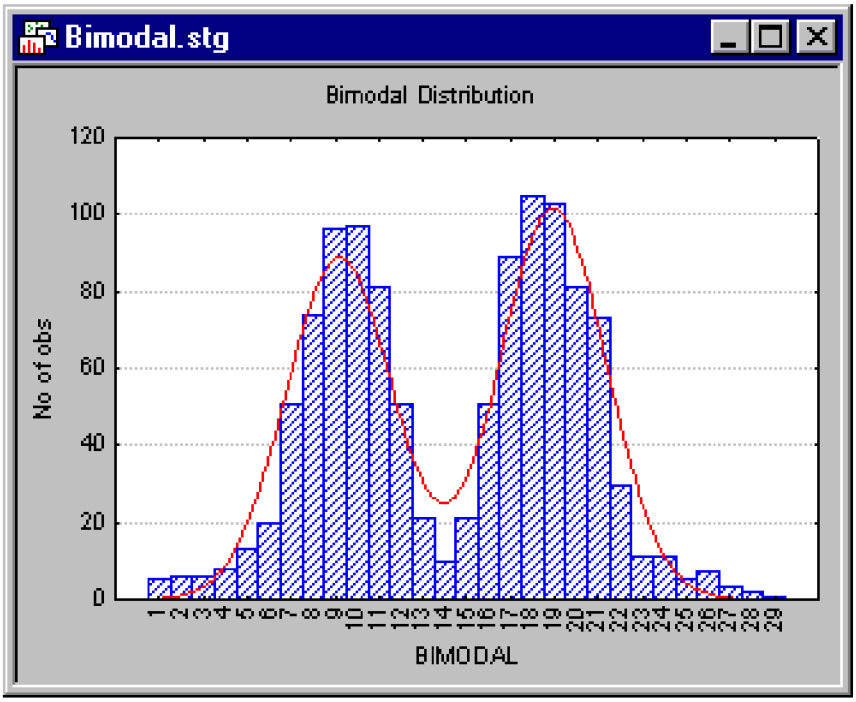


Figure 15 - Bimodal distribution histogram [Statistica help]

11 Realisation at ESIEE Paris

Expression analysis of cancer patients can further lead to an understanding of clinical outcomes. No robust genomic signature was ever found predicting the response to preoperative chemotherapy in basal-like breast cancer transcriptomic data.

How is noticed in theoretical part of this work microarray chips are used to measure the expression of tens of thousands mRNAs simultaneously. The microarray gene expression measurements are complicated and the search for potential cancer biomarkers can be improved using optimization methods. Supervised and unsupervised classifications are two methods for analyzing the gene expression of data from these devices. Supervised classification is the predictive modelling method, which give us the way how to analyze quantitative information from the datasets. It selects the representative variables from a training set (feature selection), design the classifier fit to these variables (model design) and assess the performances of prediction from similar platforms (statistical validation) or the same one (cross-validation).

The molecular signature is the set of genes whose expressions are predictive of a molecular class. From machine learning perspective, predictive models built from give signature to robustly assign the patients to their proper class across the wide of settings. Numerous methods for the gene selection have been proposed in theoretical part of this paper. The use of these classification methods is important for clinical practice because the predictor realized on smaller signature can be implemented easier and more over small signatures can have higher value for biological research.

In the present work we addressed the computing of molecular signature with use of linear discriminant analysis and bimodal index to seeking for robust classification and appropriate results.

11.1. Datasets

We assessed our predictive modeling on two datasets in oncology gain from HG133U Affymetrix microarray. Small and local tumour samples (few hundred cells) were drawn by fine needle aspiration by Lajos Pusztai from Anderson Cancer Center. Samples are from FNA biopsy, when not many cells are removed just often separated from the rest of the breast tissue so it does not show the type of cancer only if it is present or not.

Our expression data microarray contains from 22283 probe sets, indicates genes and together 138 cases. First column contains from names of n genes under study, the columns to the right contain the measurement for l healthy tissues and in complex dataset followed by m cancer tissues. So we obtained text files for samples of basal like

pathologic complete response (PCR) with 94 cases and basal like residual disease (RD) with 44 cases, HER2 samples also for PCR and RD cases and luminal-A and luminal-B BC samples. Because of the aim of the work to find gene signatures for TNBC we were mainly working with basal like samples, combine with the correction of methods on Luminal A and Luminal B, prospectively others, datasets.

Datasets are attached in the attachment of this thesis, also with programs. These datasets are loading in both case of method use in Matlab and in R.

12 Prediction using LDA

This method application was realised in space of programming language Matlab.

In each run, N=100, we are using split sample method, which contains to wrapper methods, to do random selection of $\frac{3}{4}$ cases of PCR, R' = 70 cases and of RD, N' = 33 cases as the learning sets. Rest of cases, R'' = 24 cases and N'' = 11 cases are test cases for the future predictor.

Than we are doing the random selection again on the learning cases $\frac{2}{3}$ of R', so r = 46 responders and $\frac{2}{3}$ of N', n = 22 responders. They are the learning sets of the run and the remaining cases are the test sets, r'= 24 and n'= 11 cases.

For split sample method was used randperm algorithm. It was necessary at first to modify given data and work with them without first column of names of probesets, which were saved to other file.

```
for d=1:100
   d
tic
   ss = randperm (94);
        LpCR = PCR_DataMatrix (:,ss(1:70)); % Choose of the first learning set ( 3/4
        cases of  PCR, R' = 70 cases)  from given dataset
        TpCR = PCR_DataMatrix (:,ss(71:end)); % Choose of the first test set (R'' = 24
        cases) from given dataset

        qq = randperm (44);
        Lrd = RD_DataMatrix(:,qq(1:33)); % Choose of  Learning set (RD, N' = 33
        cases)
        Trd = RD_DataMatrix(:,qq(34:end)); % Choose of  Learning set (RD, N'' = 11
        cases)
cas1 = toc;
end
```

Here is coming the question of importance of the resampling. The rest of prediction is focused on the resample cases. It seems from the figures 11 and 12 that depending on the resampling and number of iteration the p-values can be different and also for small values which are important for us.

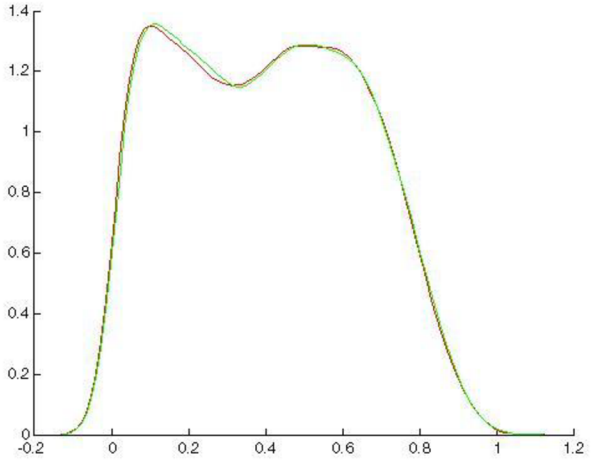


Figure 16 - one of the example of plot for the cases with (Green line) and without resampling (Red line)

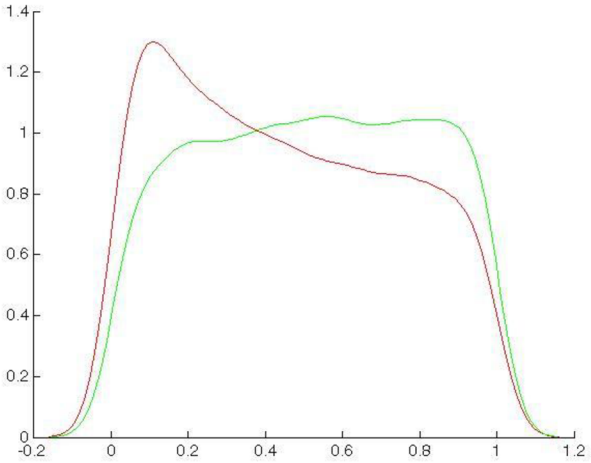


Figure 17 - other example of the cases with (Green line) and without resampling (Red line)

The first step in the predicting modelling is the selection of a molecular signature, i.e. a subset of the N variables. Because the number of subsets is 2^N, efficient methods for selecting molecular signatures are used of developed, as was mentioned in previous part of the work. Most of these methods coming from the field of Statistics as in our case used Wilcoxon-Mann-Whitney test.

For each gene of both subtypes PCR and RD is computed p-value to the Wilcoxon-Mann-Whitney test and the results are ranked by the received p-values. The low p-values show that these genes are robustly differentially expressed in the two phenotypes or classes. In the terms of optimization procedures, selecting signature, means the set of

subsets of genes, consisting in the *k* genes of smallest p-value can be placed as greedy optimization procedure. However, the "greedy" solution is not in every case optimal. The p-value is computed by function ranksum, which is Matkab function from Statistics toolbox which is returns the p-value of just a two-sided Wilcoxon rank sum test, what is synonym for Wilcoxon-Mann-Whitney test.

```
pvale = [];
for j = 1: size(LpCR,2)
   pvale = [pvale; ranksum(LpCR(:,j), Lrd(:,j))];  % Calculation of p-values for learning
   PCR and RD
end
ppvale(:,d)=pvale;
```

Subsequently, all genes are ranked in increasing order depends on their p-value, than we can select them and predict the test cases R'' and N'' by a Linear Discriminate Analysis, which, how is in more detail described in chapter 8.5., characterize or separate two or more classes of objects or events. To decide that a gene in the i(th) place shows significantly different relative expression levels with presence of cancer. The key idea of potential of biomedical genes finding can be identified as efficient solution of the LDA analysis that results representing each gene under analysis through the associated p-value, e.g. lead to set of genes with very small p-value. Prediction is by the three top ranked genes at first. For each gene, the predictor is fit on the learning sets r and n and then is compute sensitivity and specificity of the test cases prediction. The part of algorithm is shown here:

```
%% Select the 3 top ranked genes IX(1-3) and all patients for training and
% validation and put in the variables "gene_Train" and "gene_Val"
gene_Train2=DataTrain(:,IX(1:3));
gene_Val2=DataTestCompl(:,IX(1:3));
Total_Genes_3 = [Total_Genes; IX(1:3)'];

PCR_Prediction2 = classify(gene_Val2, gene_Train2, labelTrain); % Classiofication
with the first three top ranked genes, does not matter if NFIB is in or not
```

Over the runs we noticed that there is the one probe stronger than others with really low p-value compare to the rest of the top genes. It is almost every time the first top gene in three top genes, if not we also did the experiment with every iteration selection of this gene to the three top selected.

```
%% Select the top ranked gene with the smallest p-value (NFIB) + 2 top ranked genes
pos_nfib = find(IX==10875); % Is NFIB in the first 3 top genes?
if (isempty(   find(   IX(1:3) == 10875   )  )  )
   IX(3) = 10875; % If not move him to the third possition
```

end

Because we need to find how strong and robust our prediction is, the last step is to compute the specificity, sensitivity and accuracy.
The equations for the computing of the statistical measures of the performance of a binary classification test sensitivity and specificity, computed are:

$$\text{Accuracy} = \frac{Total\ Match}{Total\ number\ of\ samples}$$

$$\text{Specificity} = \frac{PCR\ Match}{Total\ number\ of\ PCR\ samples}$$

$$\text{Sensitivity} = \frac{NoPCR\ Match}{Total\ number\ of\ NoPCR\ samples}$$

12.1. Results

The first outcomes showed that the prediction were not significantly higher than random guesses but over the runs we noticed that there is the one probe stronger than others and the performances of the prediction are higher when this gene was one of the selected genes. This probe on the position 10875 in the list is indicating as probe with name 211466_at what shows on gene NFIB. The NFIB gene, encode nuclear factor I/B, is the protein-coding gene associated to lipoma of colon, polymorphous low-grade and adenocarcinoma disease. Even if it is not every time in the top three genes after direct resampling and direct top genes selection, we found it and give it as of three top genes. The performances of the prediction were higher when gene NFIB was one of the selected genes for prediction (NFIB + two other top ranked genes).

It shows the conclusion that the dataset is predictable but NFIB is necessary to this prediction because the prediction should be robust enough if random guesses are outside the 95% confidence interval of each performance criteria. Without NFIB gene none of prediction is robust so this gene is the mandatory for predicting the response. Other thing is that the confusion matrix shows that only around 30% (10 observations) are misclassified by linear discriminant function.

13 Bimodality index identification

Identification of bimodal expressed genes is an important task, as genes with bimodal expression play role in the cell differentiation, signaling and disease progression. Several algorithms have been developed to identify bimodal genes from microarray data. BI is giving, similarly as p-value, difference between outputs of two groups. The research took the look for computing Bimodality Index mainly because of bigger robustness of this method and stronger results. Research was more looking for the clear prediction of the single groups for TNBC than for predictable genes with bimodality.

At first we adopted method for computing Bimodal Index in R (Wang et al.), which shows new definition of bimodality index, from the R package from the R source "http://bioconductor.org/biocLite.R" we get straight the function BimodalIndex.

This method is based on calculation of BI for each gene from the estimate parameters of a two-component normal mixture model and then ranks the genes by BI. Advantage of this method is that BI has an intuitive interpretation because it is derived from a sample size calculation, on the other size this method is designed for microarray data under the assumption of normal distribution and have to let standard deviation same for both classes thereby can arise mistakes in final results. [44]

The bimodality BI (p) of each probe set (p) is computed together with the values l*(p) and h*(p), e.g. low and high modes of the probe sets. This computation is over a set S=R∪N of responders and non responders cases, e.g. each probe set *p* has |R|+|N| expression levels.

The bimodality index was proposed as a matrix of the bimodal feature of the gene expression levels in a set of patient cases. The higher is index the higher is the bimodal feature.

From Bimodality index computation the probe set with highest bimodality is probe set 212396_s_at, which shows on probe set KIAA0090 and is same of high bimodality for PCR, for RD is the bimodality also high, the probe is on the fifth place.

The table 1 shows the best ten genes with highest bimodality by Bimodal Index computing by Wang for complete dataset (RD+PCR), where the probe 212396_s_at is the gene of highest bimodality.

Table 1 - Bimodal index for basal_complet by Wang [50]

"203290_at"	HLA-DQA1	major histocompatibility complex
"215356_at"	TDRD12	tudor domain containing 12
"220624_s_at"	ELF5	E74-like factor 5
"201504_s_at"	TSN	Translin
"218890_x_at"	MRPL35	mitochondrial ribosomal protein
"210655_s_at"	FOXO3	forkhead box O3
"215733_x_at"	CTAG2	cancer/testis antigen 2
"209728_at"	HLA-DRB4	major histocompatibility complex
"210546_x_at"	CTAG1A/B, CTAG2	cancer/testis antigen 1A,2,1B
"212396_s_at"	KIAA0090	EMC1 ER membrane protein complex subunit 1

The histograms show the bimodality of the genes with highest bimodality of PCR, RD and complete dataset.

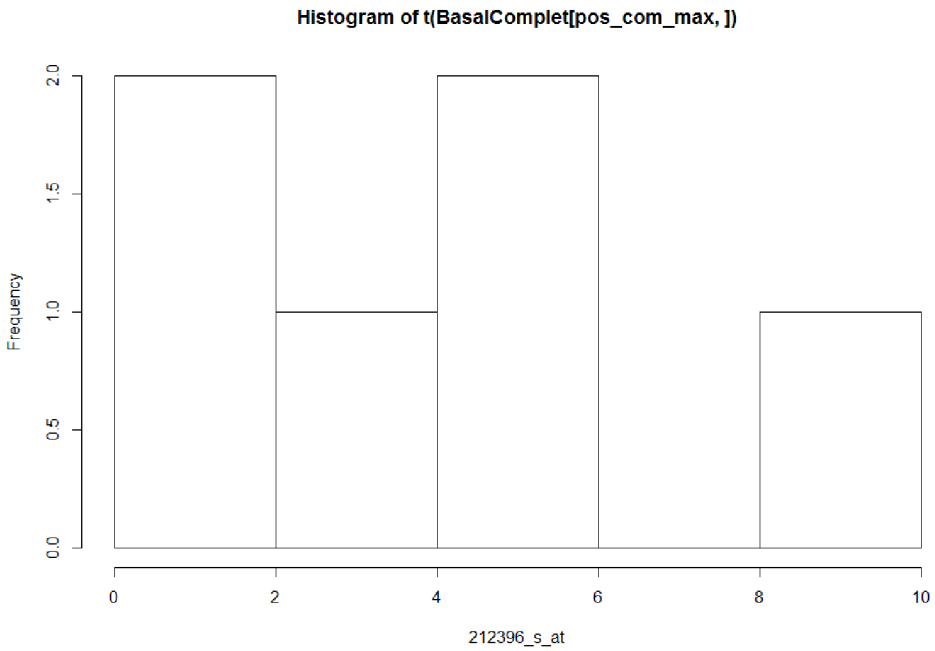


Figure 18 - gene of highest bimodality for complet basal data

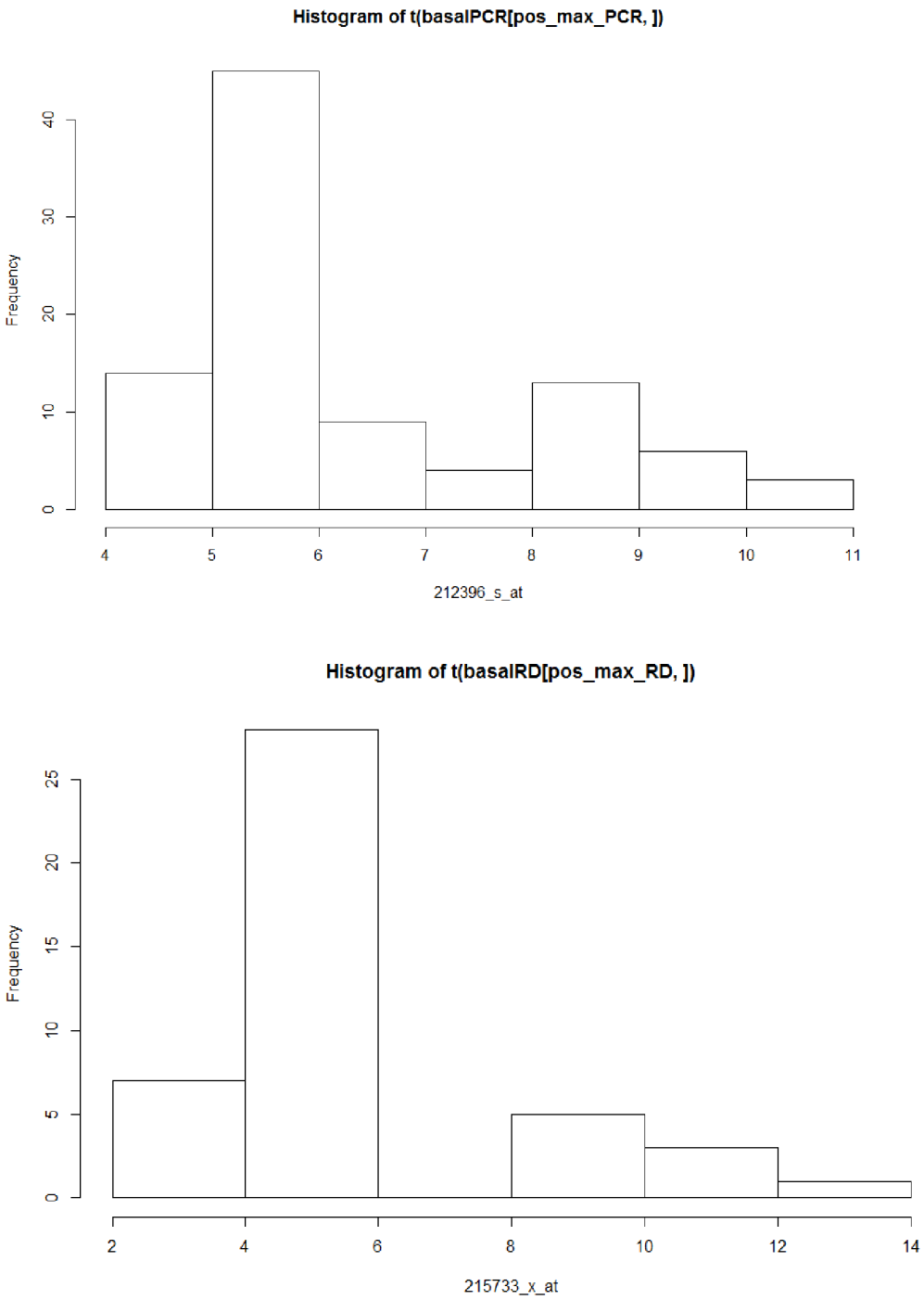


Figure 19 - Max. BI for PCR and RD of basal-like BC

Next was proposed new and more robust method depending on the quantization errors constructed in Python, Java (method defined by René Natowicz) and finally in R for computing the bimodality value of a distribution based on quantization level. The proposed method in Java is significantly the fastest than versions in Python and R. Also version in R shows little bit other results. Bimodal indexes are significantly higher,

without explanation why. The method consist from three functions, first one returns the one level quantization error:

```
qe1 <- function(E){
n <- length(E)
  m <- median(E)
  e1 = 0
h=1;while (h<=n){e1 <- sum(abs(E[h]-m));h=h+1}
  return (e1)
}
```

Second one returns two level quantization error and two quantizes istar and jstar:

```
qe2 <- function(Eprime){ # Eprime is a list of values.
  # 2 level quantization error.
  E=sort(Eprime)
  n=length(E)
  i=1; j=2;
  estar=0.0
  k=3; while(k<=n){estar=estar+(E[k]-E[j]); k=k+1}
  i=2
  while(i<=n-1){
    j=i+1
    while(j<=n){
      eij = 0.0
      k = as.integer((i+j)/2)
      l=1; while (l<=k){ eij = eij + abs(E[l]-E[i]); l=l+1}
      while (l<=n){eij = eij + abs(E[l]-E[j]); l=l+1}
      if (eij < estar){
        estar = eij
        istar=i
        jstar=j
      }
      j=j+1
    }
    i=i+1
  }
  #       cat("estar, E[i*], E[j*] : ",estar,E[istar],E[jstar])
  return( c(estar,E[istar],E[jstar])  )          # quantization error, lower mode, higher mode
}
```

Last function in this algorithm returns bimodality value:

```
bimod <- function(E){
  e1 <- qe1(E)
  # e2 <- qe2(E)
  pom <- qe2(E)
  i2 <- pom[[2]]
  j2 <- pom[[3]]
  e2 <- pom[[1]]
  bi <- (e2-e1)/e2
  return (c(bi,i2,j2))
}
```

The results of Bimodality index and new proposed algorithm show difference and we can use them for comparison of new proposed method. Wang's method shows higher bimodalities. The 212396_s_at which is rank 1 in Wang bimodality index is at rank as 104 with the definition of bimodality in new proposed method. And the gene of highest bimodality in new proposed method rating is HLA-DQA1, probeset 203290_at, histogram shown in figure 20. HLA-DQA1 is in Wang bimodality ranking on tenth place. The first ten ranked genes by highest bimodalities by new proposed method are in table 2, gene of the highest bimodality is on the last position in table 2.

Table 2 - Highest bimodalities measured by the new proposed method

216560_x_at	IGLV3-10	Immunoglobulin lambda variable 3-10
216623_x_at	TOX3	TOX high mobility box family member 3
209243_s_at	PEG3	Paternally expressed 3
203638_s_at	FGFR2	Firoblast growth factor receptor 2
211560_s_at	ALAS2	Aminlevulinate, delta-, synthase 2
220624_s_at	ELF5	E74-like factor 5
204885_s_at	MSLN	Mesothelin
205916_at	S100A7	S100 calcium binding protein A7
209728_at	HLA-DRB4	Major histocompability complex
203290_at	HLA-DQA1	Major histocompability complex

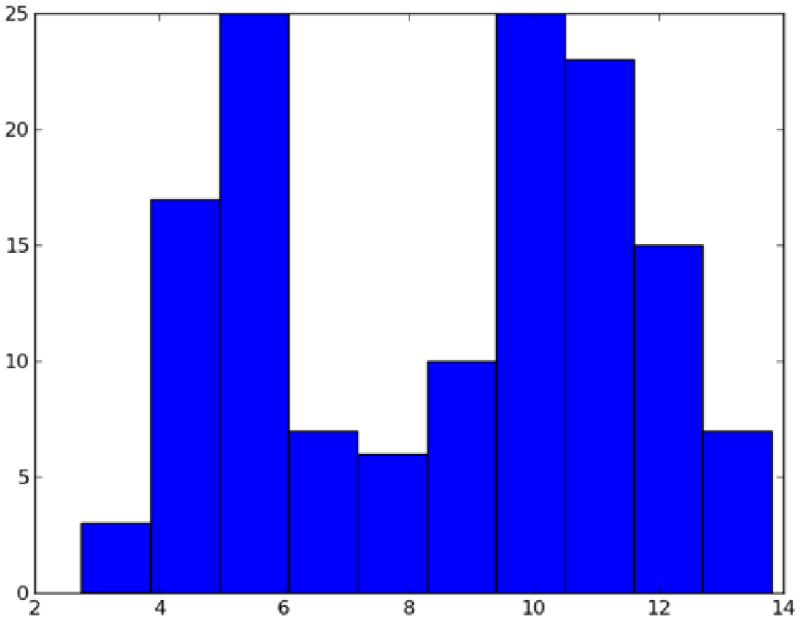


Figure 20 - The highest bimodality gene HLA-DQA1 from new proposed method

The new proposed method shows the use of advantage of bimodality index possibility to separate single classes and use them for prediction of which genes, mean to find which gene is touched by each class and select them to the right classes. We can see also from the histogram shape the possibility to differ the classes.

Resume

The first part of this work is theoretical summary of problematic around breast cancer, oncogenomics and optimization methods. The practical part connect to this theoretical introduction took place at university ESIEE Paris in terms of program ERASMUS Plus under supervisor prof. René Natowicz and in collaboration with prof. Thiago Souza from Brazil university.

The work is the only small part of several years' research in area of prediction to preoperative chemotherapy in breast cancer. The idea was application of metaheuristic methods of optimization and statistics methods to find relevant gene signatures for predicting the response to chemotherapy treatments in breast cancer.

As the most significant were identified p-values by Wilcoxon-Mann-Witney test and subsequent application of Linear Discriminant analysis to fit the predictor and predict test cases of our datasets. However, prediction accuracy does not show the good results, the method shows on the possibilities of prediction and that there is one important gene. But the method is not robust and do not give appropriate results. So because of this reason and looking for better robust and prediction next part of research is fastening to computing of bimodality index which should show significantly good results in field of gene expression analysis for finding biologically relevant data. At the end of these we are proposing methods of computing bimodality index to find biologically significant answer. The new proposed method advantage two side error shows significantly better results than Wang's method even it is also established on difference between classes near to their mean value with advantage of not necessary normal distribution and knowledge of standard deviation (compare to Wang et al. method). It goes fast in Java, mean around 25 seconds. There was made the comparison of both methods and from the results it shows that the bimodality index give the possibility to find which class of genes contain to which one, but it needs closer look and more tests, not only with the breast cancer samples but also with probably other types of tissues from other organs. Unfortunately more datasets were not available for the research. But for now it is possible to say that it is predictable and robust and can be presented in near future for significant results for clinicians.

Resources

[1] *Introduction to parallel computing*. 2nd ed. Essex: Pearson Education Limited, 2003, xx, 636 s. ISBN 02-016-4865-2.

[2] GRAMA, Ananth, Anshul GUPTA, George KARYPIS a Vipin KUMAR. *Search Algorithms for Discrete Optimization Problems* [online]. University of Denver, 2010[cit.2015-01-03]. From: http://wwwusers.cs.umn.edu/~karypis/parbook/Lectures/AG/chap11_slides.pdf. Lecture Notes. University of Denver.

[3] RÖNNQVIST, Mikael. *Discrete optimization: Exact methods* [online]. Bergen, Norway,2009[cit.2015-01-03]. From:https://www.sintef.no/globalassets/project/evitameeting/ronnqvist.pdf. Lecture Notes. Norwegian school of economics and bussiness administration.

[4] LUKE, Sean. *Essentials of metaheuristics: A Set of Undergraduated Lecture Notes*. 2. ed. lulu.com: Department of Computer Science George Mason University, 2013. ISBN 978-130-0549-628.

[5] *Greedy Introduction* [online]. 2014 [cit. 2015-01-03]. From:http://www.personal.kent.edu/~rmuhamma/Algorithms/MyAlgorithms/Greedy/greedyIntro.htm

[6] Greedy Algorithms. In: *Greedy Algorithms* [online]. [cit. 2015-01-03]. From: http://www.cs.rochester.edu/~gildea/csc282/slides/C16-greedy.pdf

[7] *Evoluční algoritmy*. Ústav biomedicínského inženýrství, 2014. Prezentace ke cvičení. VUT Brno.

[8] What is RNA. *The RNA Society* [online]. 2014 [cit. 2015-01-03]. Dostupné z: www.rnasociety.org

[9] BASSEUR, Matthieu, El-Ghazali TALBI, Antonio NEBRO a Enrique ALBA. *Metaheuristics for Multiobjective Combinatorial Optimization Problems: Review and recent issues* [online]. Unité de recherche INRIA Futurs, 2006 [cit. 2015-01-03]. From: https://hal.archives-ouvertes.fr/file/index/docid/95723/filename/RR-5978.pdf. Project. Parc Club Orsay Université.

[10] American Cancer Society. *American Cancer Society* [online]. 2014 [cit. 2015-01-03]. From: www.cancer.org

[11] Cancer Health Center. *WebMD* [online]. 2014 [cit. 2015-01-04]. From: http://www.webmd.com/cancer/

[12] BreastMark. *Medscape* [online]. 2014 [cit. 2015-01-04]. From: http://www.medscape.com/viewarticle/819503

[13] What is RNA. *Exploring life´s origins* [online]. 2014 [cit. 2015-01-04]. From:http://exploringorigins.org/rna.html

[14] SLETTO, Raymond F. a Paul E. MEEHL. *Clinical versus Statistical Prediction* [online]. 2005 [cit. 2015-01-04]. From: https://www.psych.umn.edu/faculty/grove/112clinicalversusstatisticalprediction.pdf

[15] GARDEAUX, Vincent, René NATOWICZ, WANDERLEY a Rachid CHELOUAH. *Optimization for feature selection in DNA microarrays* [online]. Paris, 2013 [cit. 2015-01-04]. Dostupné z:http://gardeux-vincent.eu/Research/NOVA13.pdf. Project. ESIEE-Paris.

[16] DUVAL, B. a J.-K. HAO. *Advances in metaheuristics for gene selection and classification of microarray data* [online]. France, 2009 [cit. 2015-01-04]. ISBN 10.1093/bib/bbp035. From:http://bib.oxfordjournals.org/content/11/1/127.long#sec-28. Project. University of Angers.

[17] AZUAJE, Francisco. *Bioinformatics and biomarker discovery: "omic" data analysis for personalised medicine*. Hoboken, NJ: John Wiley, 2010, xviii, 230 p. ISBN 978-047-0744-604.

[18] *NCBI* [online]. [cit. 2015-05-16]. Dostupné z: (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2834379/)

[19] *Microarrays in cancer research: Recent advances and future direction* [online]. [cit. 2015-05-16]. Dostupné z: http://media.affymetrix.com/support/technical/appnotes/microarrays_cancer_research_appnote.pdf

[20] *Affymetrix HG U133A* [online]. [cit. 2015-05-18]. Dostupné z: http://www.affymetrix.com/estore/catalog/131537/AFFY/Human+Genome+U133A+2.0+Array#1_1

[21] AL ANI, Tarik. DÉPARTEMENT INFORMATIQUE ET TÉLÉCOMMUNICATION, ESIEE PARIS.*Pattern recognition and machine learning: Part 1: The main components of a pattern recognition system* [online]. 2014 [cit. 2015-05-19]. Dostupné také z: http://perso.esiee.fr/~hilairex/IMC-4301B/IMEC4-2ML_RF1_last_version.pdf

[22] AL ANI, Tarik. DÉPARTEMENT INFORMATIQUE ET TÉLÉCOMMUNICATION, ESIEE PARIS.*Pattern recognition and machine learning: Part 2: Classifier* [online]. 2014 [cit. 2015-05-19]. Dostupné také z: http://perso.esiee.fr/~hilairex/IMC-4301B/IMEC4-2ML_RF2_last_version.pdf

[23] Metaheuristic. *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2015-06-02]. Dostupné z: http://en.wikipedia.org/wiki/Metaheuristic

[24] SAEYS, Y., I. INZA a P. LARRANAGA. A review of feature selection techniques in bioinformatics.*Bioinformatics*. 2007, **23**(19): 2507-2517. DOI: 10.1093/bioinformatics/btm344. ISSN 1367-4803. Dostupné také z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm344

[25] BHATTACHARYYA, Malay, Joyshree NATH a Sanghamitra BANDYOPADHYAY. MicroRNA signatures highlight new breast cancer subtypes. *Gene*. 2015, **556**(2): 192-198. DOI: 10.1016/j.gene.2014.11.053. ISSN 03781119. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0378111914013304

[26] PRAT, Aleix a Charles M. PEROU. Deconstructing the molecular portraits of breast cancer.*Molecular Oncology*. 2011, **5**(1): 5-23. DOI: 10.1016/j.molonc.2010.11.003. ISSN 15747891. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1574789110001274

[27] SABATIER, Renaud, Pascal FINETTI, Arnaud GUILLE, José ADELAIDE, Max CHAFFANET, Patrice VIENS, Daniel BIRNBAUM a François BERTUCCI. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Molecular Cancer*. 2014, **13**(1): 228-. DOI: 10.1186/1476-4598-13-228. ISSN 1476-4598. Dostupné z: http://www.molecular-cancer.com/content/13/1/228

[28] PRAT, A, A LLUCH, J ALBANELL, W T BARRY, C FAN, J I CHACÓN, J S PARKER, L CALVO, A PLAZAOLA, et al. Predicting response and survival in chemotherapy-treated triple-negative breast cancer. *British Journal of Cancer*. 2014, **111**(8): 1532-1541. DOI: 10.1038/bjc.2014.444. ISSN 0007-0920. Dostupné z: http://www.nature.com/doifinder/10.1038/bjc.2014.444

[29] SIMON, Richard. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *ACM SIGKDD Explorations Newsletter*. 2003, **5**(2). DOI: 10.1145/980972.980978.

[30] KOBOLDT, Daniel C., Robert S. FULTON, Michael D. MCLELLAN, Heather SCHMIDT, Joelle KALICKI-VEIZER, Joshua F. MCMICHAEL, Lucinda L. FULTON, David J. DOOLING, Li DING, et al. Comprehensive molecular portraits of human breast cancer. *Nature*. 2012-9-23, **490**(7418). DOI: 10.1038/nature11412. ISSN 0028-0836.

[31] SIMON, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*. 2003, **89**(9): 1599-1604. DOI: 10.1038/sj.bjc.6601326. ISSN 0007-0920. Dostupné z: http://www.nature.com/doifinder/10.1038/sj.bjc.6601326

[32] DUPUY, A. a R. M. SIMON. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI Journal of the National Cancer Institute*. 2007, **99**(2): 147-157. DOI: 10.1093/jnci/djk018. ISSN 0027-8874. Dostupné z: http://jnci.oxfordjournals.org/cgi/doi/10.1093/jnci/djk018

[33] GYŐRFFY, Balázs, Thomas KARN, Zsófia SZTUPINSZKI, Boglárka WELTZ, Volkmar MÜLLER a Lajos PUSZTAI. Dynamic classification using case-specific

training cohorts outperforms static gene expression signatures in breast cancer.*International Journal of Cancer*. 2015, **136**(9): 2091-2098. DOI: 10.1002/ijc.29247. ISSN 00207136. Dostupné z: http://doi.wiley.com/10.1002/ijc.29247

[34] VAZQUEZ, Miguel, Victor DE LA TORRE, Alfonso VALENCIA, Fran LEWITTER a Maricel KANN. Chapter 14: Cancer Genome Analysis. *PLoS Computational Biology*. 2012, **8**(12): e1002824-. DOI: 10.1371/journal.pcbi.1002824. ISSN 1553-7358. Dostupné z: http://dx.plos.org/10.1371/journal.pcbi.1002824

[35] FAY, Michael P. a Michael A. PROSCHAN. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*. 2010, **4**: 1-39. DOI: 10.1214/09-SS051. ISSN 1935-7516. Dostupné z: http://projecteuclid.org/euclid.ssu/1266847666

[36] *Statistics solutions* [online]. 2015 [cit. 2015-06-13]. Dostupné z: http://www.statisticssolutions.com/

[37] DE NEVE, Jan, Olivier THAS, Jean-Pierre OTTOY a Lieven CLEMENT. An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology*. 2013,**12**(3): -. DOI: 10.1515/sagmb-2012-0003. ISSN 1544-6115. Dostupné z: http://www.degruyter.com/view/j/sagmb.2013.12.issue-3/sagmb-2012-0003/sagmb-2012-0003.xml

[38] BOULESTEIX, A.-L. WilcoxCV: an R package for fast variable selection in cross-validation.*Bioinformatics*. 2007, **23**(13): 1702-1704. DOI: 10.1093/bioinformatics/btm162.

[39] *Coursera* [online]. 2015 [cit. 2015-06-13]. Dostupné z: https://www.coursera.org/

[40] WANG, Jing, Sijin WEN, W. Fraser SYMMANS, Lajos PUSZTAI a Kevin R. COOMBES. The Bimodality Index: A criterion for Discovering and ranking bimodal signatures from Cancer gene expression profiling data. *Cancer informatics*. 2009, (7). Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2730180/

[41] ERTEL, Adam. Bimodal gene expression and biomarker discovery. *Cancer Informatics*. 2010, (9:11-14). Dostupné z: http://www.ncbi.nlm.nih.gov/pubmed/20234772

[42] HELLWIG, Birte, Jan G HENGSTLER, Marcus SCHMIDT, Mathias C GEHRMANN, Wiebke SCHORMANN a Jorg RAHNENFUHRER. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics*. 2010, **11**(1): 276-. DOI: 10.1186/1471-2105-11-276. ISSN 1471-2105. Dostupné z: http://www.biomedcentral.com/1471-2105/11/276

[43] HAURY, Anne-Claire, Pierre GESTRAUD, Jean-Philippe VERT a Muy-Teck TEH. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE*. 2011, **6**(12): e28210-. DOI: 10.1371/journal.pone.0028210. ISSN 1932-6203. Dostupné z: http://dx.plos.org/10.1371/journal.pone.0028210

[44] TONG, P., Y. CHEN, X. SU a K. R. COOMBES. SIBER: systematic identification of bimodally expressed genes using RNAseq data. *Bioinformatics*. 2013, **29**(5): 605-613. DOI: 10.1093/bioinformatics/bts713. ISSN 1367-4803. Dostupné z: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts713

[45] Cancer research UK. *Cancer research UK* [online]. 2013 [cit. 2015-06-17]. Dostupné z: http://www.cancerresearchuk.org/

[46] *Youtube* [online]. 2005 [cit. 2015-06-17]. Dostupné z: https://www.youtube.com/

[47] CureJoy. *CureJoy* [online]. 2013 [cit. 2015-06-17]. Dostupné z: http://www.curejoy.com/

[48] *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2015-06-17].

[49] SÁNCHEZ-PEÑA, Matilde L., Clara E. ISAZA, Jaileene PÉREZ-MORALES, Cristina RODRÍGUEZ-PADILLA, José M. CASTRO a Mauricio CABRERA-RÍOS. Identification of potential biomarkers from microarray experiments using multiple criteria optimization. *Cancer Medicine*. 2013, **2**(2): 253-265. DOI: 10.1002/cam4.69. ISSN 20457634. Dostupné z: http://doi.wiley.com/10.1002/cam4.69

[50] *Genecards* [online]. 1996 [cit. 2015-06-18]. Dostupné z: www.genecards.org

Shortcuts

OP	Optimization problem
MOPs	Multi Objective Problems
MCOPs	Multi Objective Combinatorial Optimization Problems
DOP	Discrete Optimization Problem
EC	Evolutionary Computation
EA	Evolutionary Algorithm
GA	Genetic Algorithm
ES	Evolutionary Strategies
DNA	Deoxyribonucleic Acid
IonOGen	Interactive Oncogenomic Database
RTCGD	Retrovirus Tagged Cancer Gene Database
RNA	Ribonucleic Acid
ER	Estrogen Receptor
PR	Progesteron Receptor
HER2	Human Epidermal growth factor Receptor 2
ML	Machine Learning
LDA	Linear Discriminant Analysis
FS	Feature Selection Technique
BC	Breast Cancer
TNBC	Triple Negative Breast Cancer
HG	Human Genome
PCR	Pathologic complete Response (responders)
RD	Residual Disease (Non responders)
Pdf	Probability density function
BLBC	Basal Like Breast Cancer
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
mRNA	messenger RNA
tRNA	transfer RNA
rRNA	ribosomal RNA
cDNA	complementary DNA
RT-PCR	Real Time Polymerase Chain Reaction
SAGE	Serial analysis of gene expression
DOP	Discrete Optimization Problem
WMW	Wilcoxon-Mann-Whitney test