Oligonucleotide microarrays in biomedical sciences – the use and data analysis

Robert Milewski[1], **Anna Justyna Milewska**[1], **Jan Czerniecki**[2,3]

[1] Department of Statistics and Medical Informatics, Medical University of Bialystok
[2] Department of Biology and Pathology of Human Reproduction, Institute of Animal Reproduction and Food Research of Polish Academy of Sciences in Olsztyn
[3] Department of Cytobiochemistry, Institute of Biology, University of Bialystok

Abstract. The methods used in biomedical research are becoming inadequate to meet current challenges. Frequently occurring problem is the need to find the differentiation tests according to phenotypic features or the particular phenomenon. Previously used morphological evaluation or other laboratory tests many times do not allow for adequate determination of differentiating attributes. In recent years there has been considerable scientific and technological progress in the fields such as genomics, transcriptomics, proteomics and metabolomics, which allow to move the search area into the molecular level. It allows the use of advanced molecular techniques such as PCR or oligonucleotide microarrays and thus allows to compare the gene expression profiles of different types of cells and tissues. The microarray experiment data allow to determine the correlation between the expression of selected genes or even entire genotypes of the phenotypic features, characterizing the studied group. The collected data can not be analyzed using traditional statistical methods, since the number of cases is much higher than the number of considered attributes. For this reason, new statistical methods and procedures are used for microarray data analysis which may focus on theoretical or practical aspects. Theoretical aspect is related to the selection of specific genes expression, finding the ontology or metabolic pathways that are associated with the analyzed phenomenon. The practical aspect can be the creation of a predictive model that can allow to predict the specific phenonon occurrence in the future during the studies of new patients. Microarray experiments and analysis of the obtained results begin new chapters of particular phenomena investigation, which is another big boost in the biomedical sciences development.

Introduction

It is a common practice in the biomedical sciences making the differentiation of the tested material (cells, tissues) in order to find specific phenotypic features or the particular phenomenon (e.g. illness). These studies have been carried out using morphological assessment or other laboratory tests, which do not always allow for adequate determination of specific characteristic.

In recent years there has been a significant progress in such areas as genomics, transcriptomics, proteomics and metabolomics, which allow to

ISBN 978–83–7431–296–7 ISSN 0860-150X

move the searching area into the molecular level. Genomics is the analysis of the tested organisms genome, in order to determine the genetic material sequence, genome mapping and estimate the relationships and interactions within it.

Transcriptomics is the study of genome activity (individual genes expression) by determining the transcriptome changes dependent on place and time. The transcriptome is a set of mRNA molecules, resulting in a particular cell during particular time as a result of expression of specific genes. The study using transcriptomes methods enable to detect and identify a large number of different RNA molecules. Among these methods, particular attention should be paid to the microarray analyses and their possibilities to study of the entire genomes activity.

Proteomics, in turn, is the study of protein synthesis and structure, relationships between them, as well as their function examination. This area refers to research carried out on a large scale, concerning the entire proteomes. The complexity of the proteins structure as well as their number and variability mean that proteomics is a much more complicated science than genomics and transcriptomics. The protein microarrays, whose principles are similar to DNA microarrays are widely used in the determination of proteomic profiles or an interaction between protein.

The fourth discipline, recently developed is metabolomics. It concerns the study of all metabolites set of the whole body, particular tissue or cell (metabolome). Despite the great progress that has been made in recent years, metabolomics is not understood and developed as well as the previously described "omics" sciences. Partially due to its character and complexity.

Among a variety of techniques including the aforementioned four areas of molecular biology the microarray analysis is one of the most versatile techniques in biomedical sciences. It allows to specify and compare the gene expression profile of different types of cells or tissues. The collected data allow to determine the correlation between the expression of selected genes (or even entire genomes) and the phenotypic, characteristic of the studied groups. This allows for conducting experiments in virtually every field of medicine, including areas where using traditional methods has already been considered exhausted.

Microarrays

The first DNA microarrays were produced in the early 90s of the twentieth century to serve as a tool for DNA sequencing, mutations recovering

and mapping of genomes [11–12]. They became very popular when it was discovered that they can be used to study gene expression profiles [23].

DNA Microarray is a collection of molecular probes which adhere to the base (usually glass or silicon) in a particular order. Due to the construction of probes we can distinguish two types of microarrays: cDNA – long probe, even full-sequences corresponding to the mRNA, and oligonucleotide microarrays – short probes, typically 25–70 nucleotides. The latter can produce a much denser packing of probes on the base, mainly because they are synthesized in situ and not imprinted. Synthesis in situ adds nucleotides one by one until the end of the growing oligonucleotide [8]. To get required complex oligonucleotide sequences photolithography is used. Certain areas of the base are exposed to light, causing their activation and then the selected type of nucleotide is applied on the plate. Nucleotides bind only to chains that had previously been subjected to activating light, and thus each of oligonucleotides is elongated by one (specific) nucleotide. Nucleotide chains of required length and the sequences in the certain positions are obtained by repeating of this process.

Currently oligonucleotide microarrays can contain up to 300 000 probes per square centimeter which gives more than one million probes on a single plate. Gene expression analysis usually do not require such amounts of probes. Therefore more microarrays are placed on one plate what allows to reduce the cost of the experiment. These parameters are determined depending on the nature of the experiment and the type of tested material. Currently there are plates which allow to analyze the expression of the entire human genome for eight samples on a singular plate. To fully reflect the complexity of the microarray structure it is also worth noting that a singular probe on the plate is composed of about one billion copies of the same oligonucleotide chains [8].

The structure and microarray technologies differ, depending on the companies involved in preparation of these products. Nowadays the most commonly used microarrays in Poland are prepared by Agilent and Affymetrix companies [Fig. 1].

Each microarray experiment consists of several basic steps [20]. These include, among others:

- RNA isolation
- labeling
- hybridization
- scanning
- image analysis
- statistical data analysis

Fig. 1. Microarray chips made by Affymetrix company

Depending on the type of experiment one or more groups are analyzed. Often, as in classical biomedical experiments, the test and control groups are analyzed. For example, the experiments in oncology are performed using tumor tissue from treated patients as a test group, while a control group is healthy tissue preparations.

Depending on the character of the experiment as well as the plate type one- or two-colored labeling can be used. Using one color labeling each of the identical samples stained with the same dye and hybridize to different microarrays [23]. Two-color labeling is a labeling of two different samples with two different fluorescent dyes (e.g. Cy3 and Cy5) and then combining of these two preparations and hybridization with one microarray. Two colors labeling can be used for direct comparison of the test and the control group or for simultaneous comparison each of the preparations with specially prepared background, which often allows for better results.

A key step of the experiment is the hybridization plate. It is possible due to the complementarity of nucleic acids and occurs when oligonucleotide chains on the plate have complementary sequences to the nucleic acid coming from the tested preparation. It takes place under particular conditions in the hybridization chamber and to ensure the ability to connect almost all the complementary sequences usually lasts overnight.

Then the plate after hybridization is scanned using confocal laser, resulting in the image where the colored points correspond to the intensity of the

signal and thereby the level of expression of a genes in the tested samples [Fig. 2]. Therefore the resulting image is analyzed in order to save the color and intensity for individual spots using numerical values. Obtained in this way data is subjected to quality controls and a complex statistical analysis.

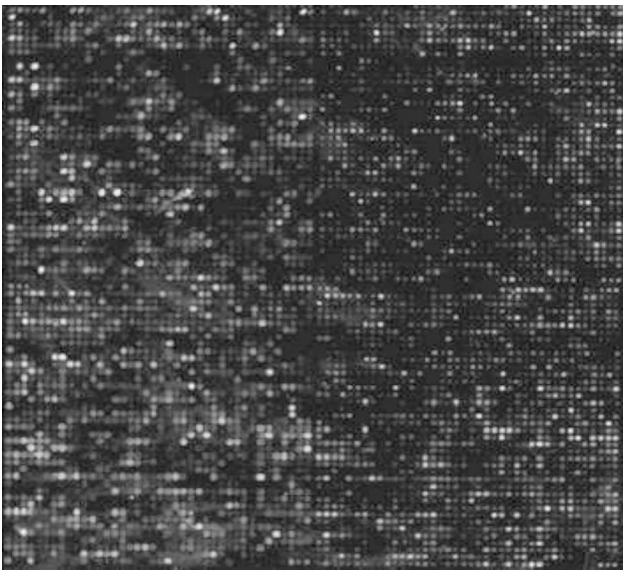

Fig. 2. The image of microarray scanned with confocal laser

Microarrays data

Recorded information in digital form, depending on the technology used may contain various parameters, but the most commonly used in the further analysis are [2]:

- signal intensity for each of the spots with the coordinates of the spot,
- background intensity for each of the spots – this is the average intensity of pixels surrounding the spot, taken as its background,
- pixel intensity distribution – areas of increased and decreased intensity may suggest the occurrence of disturbances during the experiment, in this case obtained data may require appropriate correction,
- spot morphology – the shape and size of individual spots can provide information about the quality of represented data.

The format of ultimately saved for statistical analysis data depends on the type of software. Often on the basis of the collected data sets the intensity of the signal for each spot and for each sample is measured and such infor-

mation is recorded in the form of a matrix where the rows correspond to spots and columns – samples. These arrays typically have a size of several or hundreds of columns and several thousands of rows.

Analysis of microarrays data

The collected data can not be analyzed with traditional statistical methods, since the number of analyzed cases (spots, genes) is much higher (typically thousands) of the number of considered attributes (elements of the population – typically several tens of units). In addition to traditional statistical methods, a number of advanced methods were developed and applied in medicine (eg. neural networks). They are generally called data mining methods [17–18]. However, the nature of the microarrays data necessitated the development of completely new methods and statistical procedures dedicated to these data analysis. They resulted in the development of theories of concomitant test of multiple hypotheses, based on measurening of error such as false discovery proportion (FDP) or proposed by Benjamin and Hochberg – false discovery rate (FDR) [7, 10]. The comprehensive studies, showing how to comprehensively carry out the process of microarray data analysis were formed [2].

Of course there is no possibility to analyze of this type of data without computers and statistical programs. There are numerous different kinds of programs and statistical software on the market, alike more universal and specially dedicated to microarrays, which greatly facilitate the analysis performance. Companies producing equipment and reagents for microarray experiments often develop and distribute specialized software designed to analyze the results (e.g. GeneSpring GX software). These programs are highly automated and do not require particular knowledge of advanced mathematics and statistics. However, on the other hand, they impose certain patterns of conduct, what can be a limitation for advanced biostatisticians and bioinformaticians. Therefore, they choose the software more flexible and more versatile (e.g. Stata, SAS Microarray or R software) [15].

There is no a scheme of microarray data analysis. It depends on the type of experiment as well as the model chosen by the researchers [1]. However, there are a few basic steps that should be included in any analysis: the quality control, various types of data normalization, filtering spots, extract a set of genes for which statistically significant differences were observed in expression between the analyzed groups. [Fig. 3] is a graph of the distribution of the intensity of expression for several samples after normalization with GeneSpring GX software.

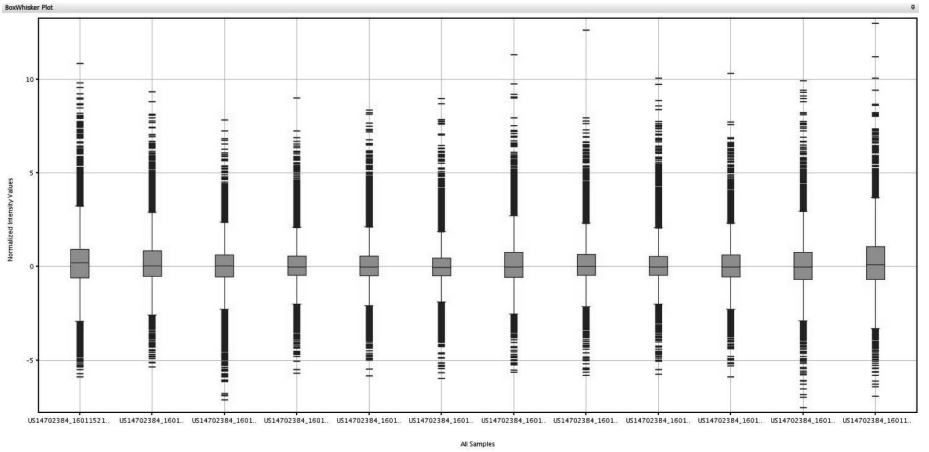

Fig. 3. Graph of the distribution of expression intensity after normalization

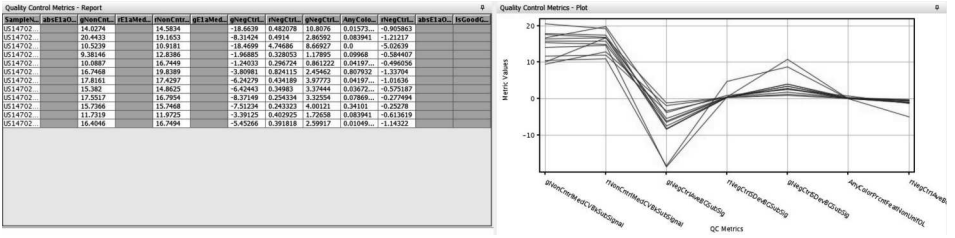

Fig. 4. Quality control metrics (QCM)

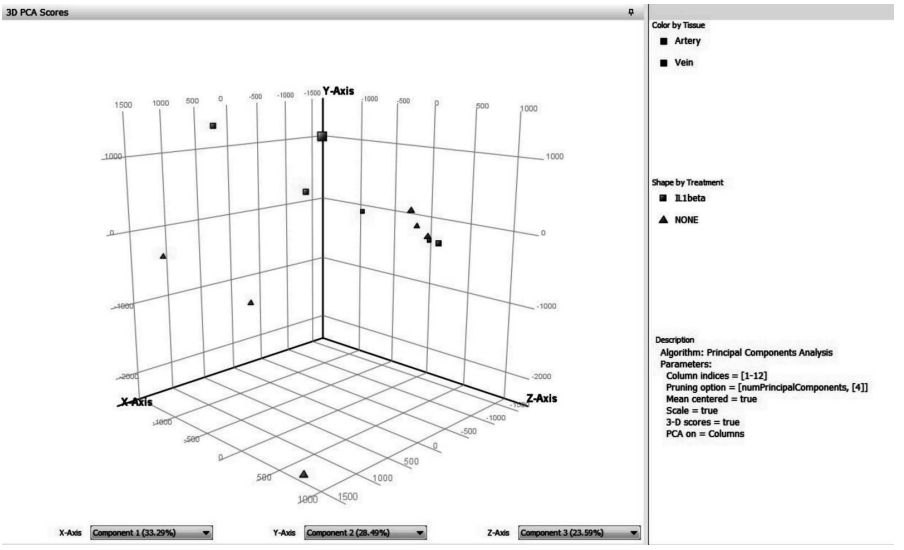

Fig. 5. Results of principal component analysis (PCA)

This application allows at this stage of the analysis to include determination of quality control metrics (QCM) [Fig. 4], to make a principal component analysis (PCA) [Fig. 5] and analyze statistical significance of differences in gene expression – analysis of variance test with multiple comparisons correction Benjamin–Hochberg [7, 10] and a fold-change analysis.

GeneSpring GX software application also provides the opportunity to illustrate subsequent stages of analysis results using various types of tables and graphs (e.g. Venn diagrams, dendrograms).

Determination of a set of genes which expression differentiates the occurrence of the examined phenomenon is a characteristic point opening following stages of statistical analysis and interpretation of molecular research. In the classical approach (without prediction) the ontology of genes, associated with a previously selected set of genes, as well as metabolic pathways in which these genes are involved are looking at the later stage. Ontologies describing the properties of genes and their gene products, consist of three types corresponding to the basic biological research areas:

- molecular function of gene products,
- their role in the multistep biological processes,
- their physical structure – as a components of the cell.

Metabolic pathways are series of successive biochemical reactions where one reaction product is a substrate for another. Ontology or the metabolic pathway are based on selected genes are a structure which action has also impact on the analyzed phenomenon and often allows to find hitherto unknown mechanisms involved in the investigated process.

A powerful research tool to facilitate finding such ontology, metabolic pathways or genes clusters functionally correlated is available on the web application Database for Annotation, Visualization and Integrated Discovery (DAVID).

Predictive model

In the clinical sciences predictive model is used more often than the classical model of microarray data analysis. It consists of creating a model, based on the learning data sets, which allow to predict the condition of the analyzed phenomena for subsequent cases. For example, based on survival time the patients undergoing a medical treatment, it is possible to create a model that will predict the estimated time of survival of another patients undergoing the same treatment.

Microarray analysis allows for the creation of such models where the data set refers to the expression of a set of genes that differentiate the occurrence of the studied phenomenon. With this approach, the less important are the mechanisms responsible for differences in gene expression but more important is the ability to predict the phenomenon of new patients and the application of the model in clinical practice. You can even produce a small, dedicated microarray that they examine only a selected small set of genes responsible for the occurrence of the studied phenomenon [13].

The procedure creation of predictive model is similar to the classical model until selection the set of genes whose expression differs between patients in two groups of interest. The information recorded at this stage for further analysis is a combination of the three numbers for each sample and for each gene:

- the logarithm of expression intensity (geometric mean of intensity of both channels)
- the logarithm of the expression factor
- the level of significance for the value of the logarithm of expression.

Such a model was created during the examination of survival of patients with diagnosed breast cancer [21–22]. The study involved women diagnosed with a tumor smaller than 5 cm, class T1 or T2, in which there was no lymph node metastasis (N0) whose age at diagnosis moment did not exceed 55 years and there was no previous history of cancer. Patients were diagnosed between 1983–1996. All patients had a modified radical mastectomy or breast conserving surgery, then they were monitored until death or until the examination in the case of living patients. Two dependent variables were fixed: the occurrence of distant metastases and death due to cancer. Analysis were performed for each of them separately. Women were divided into two groups:

- the group of "promising" patients without distant metastases (or death in the case of the second variable) within 5 years after the diagnosis of cancer,
- the "negative prognosis" where distant metastasis developed (or death occurred) before the end of 5 years period from the cancer diagnosis.

For both groups of patients the microarray expression analysis were carried out using tumor tissue frozen at the time of diagnosis.

For the final analysis 77 patients (44 without metastases and 33 with metastases after 5 years) and 24 483 genes were qualified. Then the first selection of genes using T-test was performed to determine differences in the expression profile between the groups. Seventy genes which strongly differen-

tiate the two examined groups were selected. For those genes a 70-dimensional vector of averaged expression for the group of patients with good prognosis was assigned. Then the correlation between each patient (in both groups) and the average expression vector (for this purpose is well suited cosine correlation based on a scalar product – the cosine of the angle between vectors) was determined. The correlation determined in this way proved to be a good predictor of distant metastases.

The created model has been tested on a group of validation using the leave-one-out crossvalidation. To confirm the differences the Kaplan-Meier analysis was performed. In subsequent stages of the study the test group was extended with patients with metastatic nodes (N1) and the independence of the model on factors that potentially could affect the results was demonstrated. These factors are namely:

- node metastases occurrence,
- the center of data origin,
- storage time of frozen tissue.

A small microarray for clinical applications (MammaPrint microarray) was prepared based on the carried out experiments. Then the identity of results obtained using commercial and prepared microarrays was demonstrated [9, 13]. The main goal of clinical application of small microarray is less severe treatment of patients who end up in promising groups (e.g. no need for radiotherapy and chemotherapy).

A similar experiment was carried out by a team dealing with treatment of non-small cell lung cancer (NSCLC) [19]. A predictive model of distant metastases was created based on selected during the microarray experiment 72 genes.

Moreover, the attempts of creation of predictive model of the reproductive potential of oocytes and effectiveness of infertility treatment based on cumulus cells gene expression were performed by a team of prof. Samir Hamamah [4–6, 14]. There is no possibility to use the oocyte material for molecular research because it would lead to its damage and the inability to get an embryo and eventual pregnancy. However, it has been shown that cumulus cells have the impact on the quality of the oocyte. They are responsible for oocyte nourishment. Therefore, they were used as a source of material for microarray studies. It was established that there are some genes whose expression differentiates the oocyte with positive outcomes of in vitro fertilization and oocyte taken from patients with IVF failure [3, 16]. It allows to design a small microarray which will support the selection of oocyte with the greatest fertility potential.

Conclusions

Microarray analysis of gene expression brings new information and begins a new chapters of analysis different phenomena, especially in the biomedical sciences. Availability of appropriate equipment and technology at the moment is not large, probably due to high costs. However, this technique is successively applied in medicine, bringing new opportunities and throwing new light on the phenomenon, where traditional methods seem to be already exploited.

Microarrays experiments have forced the development of special statistical methods for data analyzing due to the reverse proportion between the number of attributes and cases. The ability of gene expression determination combined with advanced statistical methods for data analysis is a powerful method in biomedical sciences which will be constantly developed in the future and will undoubtedly bring many revolutionary discoveries in medicine.

REFERENCES

[1] Allison D. B., Cui X., Page G. P., et al., Microarray data analysis: from disarray to consolidation and consensus, Nature Reviews Genetics, 7 (1), pp. 55–65, 2006.

[2] Amaratunga D., Cabrera J., Exploration and analysis of DNA microarray and protein array data, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008.

[3] Anderson R. A., Sciorio R., Kinnell H., et al., Cumulus gene expression as a predictor of human oocyte fertilisation, embryo development and competence to establish a pregnancy, Reproduction, 138, pp. 629–637, 2009.

[4] Assou S., Anahory T., Pantesco V., et al., The human cumulus-oocyte complex gene-expression profile, Human Reproduction, 21 (7), pp. 1705–1719, 2006.

[5] Assou S., Haouzi D., Mahmoud K., et al., A non-invasive test for assessing embryo potential by gene expression profiles of human cumulus cells: a proof of concept study, Molecular Human Reproduction, 14, pp. 711–719, 2008.

[6] Assou S., Haouzi D., De Vos J., et al., Human cumulus cells as biomarkers for embryo and pregnancy outcomes, Molecular Human Reproduction, 16, pp. 531–538, 2010.

[7] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Statist. Soc. Ser. B 57, pp. 289–300, 1995.

[8] Brown T. A., Genomy, PWN, 2009.

[9] Bueno-de-Mesquita J.M., van Harten W. H., Retel V. P., et al., Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER), Lancet Oncology, 8 (12), pp. 1079–1087, 2007.

[10] Dudziński M., Furmańczyk K., Procedury jednoczesnego testowania wielu hipotez i ich zastosowania w analizie mikromacierzy DNA, Matematyka Stosowana, 8, pp. 84–108, 2007.
[11] Fodor S. P., Rava R. P., Huang X. C. et al., Multiplexed biochemical assays with biological chips, Nature, 364, pp. 555–556, 1993.
[12] Fodor S. P., Read J. L., Pirrung M. C. et al., Light-directed, spatially addressable parallel chemical synthesis, Science, 251, pp. 767–773, 1991.
[13] Glas A. M., Floore A., Delahaye L. J., et al., Converting a breast cancer microarray signature into a high-throughput diagnostic test, BMC Genomics, 7, pp. 278–287, 2006.
[14] Hamamah S., Fallet C., Gene expression profile of human cumulus cells: clinical applications for IVF, J Gynecol Obstet Biol Reprod (Paris), 1 Suppl., pp. 5–7, 2010.
[15] Maciejewski H., Konarski Ł., Jasińska A., et al., Analysis of DNA microarray data, methods and tools, Bio-Algorithms and Med-Systems, 1 (1/2), pp. 129–132, 2005.
[16] McKenzie L. J., Pangas S. A., Carson S. A., et al., Human cumulus granulose cell gene expression: a predictor of fertilization and embryo selection in women undergoing IVF, Human Reproduction, 19, pp. 2869–2874, 2004.
[17] Milewski R., Jamiołkowski J., Milewska A. J., et al., Prognosis of the IVF ICSI/ET procedure efficiency with the use of artificial neural networks among patients of the Department of Reproduction and Gynecological Endocrinology, Ginekologia Polska, 80 (12), pp. 900–906, 2009.
[18] Milewski R., Malinowski P., Milewska A. J., et al., The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 35–46, 2010.
[19] Roepman P., Jassem J., Smit E. F., et al., An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer, Clinical Cancer Research, 15 (1), pp. 284–290, 2009.
[20] Stępniak P., Handschuh L., Figlerowicz M., DNA microarray data analysis, Biotechnologia, 4 (83), pp. 68–87, 2008.
[21] van't Veer L. J., Dai H., van de Vijver M. J., et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, pp. 530–536, 2002.
[22] van de Vijver M. J., He Y. D., van't Veer L. J., et al., A gene-expression signature as a predictor of survival in breast cancer, New England Journal of Medicine, 347, pp. 1999–2009, 2002.
[23] Żmieńko A., Handschuh L., Góralski M., DNA microarrays in structural and functional genomics, Biotechnologia, 4 (83), pp. 39–53, 2008.