Polytechnic University of Bucharest

Faculty of Computer Science

Department of Automation and Systems Engineering

Thesis

Molecular pathogenesis analysis of breast cancer using integrated microarray analysis and gene modeling

PhD supervisor:

Prof.dr.ing. Cătălin Buiu

PhD student:

Ing. Irina-Oana Lixandru-Petre

Bucharest, 2022

Contents of the doctoral thesis summary

Contents of the doctoral thesis

Keywords

Genes ; microarray ; bayesian networks ; cancer ; mutations ;

Chapter 1. Introduction

Bioinformatics is the science at the intersection of various fields such as biology, medicine, mathematics or computer science. Defined as the discipline that refers to the analysis and interpretation of biological data in order to obtain information and knowledge, this branch aims to reach a deeper understanding of the mechanisms of life, at any level.

Gene biological systems are characterized by continuous communication relationships with other cells or genes, alterations in this complex communication system can have serious consequences on cellular stability, increasing the risk of promoting dangerous diseases on the health of organisms. These diseases include cancer, a change in cell balance, characterized by a high number of altered cells, which divide indefinitely and form tumors that can invade the whole body.

This paper addresses microarray analysis, through an integrated technique of interpreting gene expressions from modified breast tissue, in order to better understand how mutations can be transmitted from one gene to another, leading to cell formation. cancers that damage the whole body.

Cancer is and will remain a global problem of altering the health of individuals. Although a high level of knowledge of information has been reached, specialists have not yet been able to find an answer to all the processes and modes of communication that take place before a cell undergoes so many and different mutations that it gives rise to clones that promote the formation of neoplastic tissues. The motivation for choosing this topic is therefore the major implication that cancer has in the public health system and how it acts from a micro level, of a cell mutation, to a macro level, of the formation of metastases, managing to it successfully overcomes the barriers that the body raises in order to maintain cellular stability.

The thesis is structured in seven chapters, followed by an appendix that includes the code from the R programming language and the bibliography part, which was the basis of the theoretical information in this paper.

Chapter 2 includes the research context and the field of applicability, representing an introduction in the fields of bioinformatics and data mining. It begins with the presentation of basic biological concepts related to genetic material and the regulation of gene expression, continuing with the presentation of the main existing biological databases and formats. In the following, the principles of two commonly used microarray technologies, namely those based on complementary deoxyribonucleic acid and oligonucleotides, are presented in detail, the chapter concluding with the presentation of several computational challenges related to the analysis of gene expressions.

Chapter 3 includes a presentation of the existing framework related to the analysis of biological processes in various forms. Subchapters on the mathematical modeling of biological processes are discussed, continuing with intelligent techniques that emulate functions specific to human beings or other biological systems (fuzzy, neural networks, genetic algorithms) and information related to gene expression alignments. The chapter concludes with a description of multiple gene expression analyzes, including those based on the Bayesian probabilistic model, Bayesian networks, the hidden Markov model, the least squares method, principal component analysis, or hierarchical grouping.

Chapter 4 discusses microarray analysis in detail, and what are the main steps in such a genetic analysis. We are talking about the data acquisition part, the data preprocessing and reduction part, by eliminating redundant attributes, continuing with the presentation of several methods of grouping and investigating the differential expression by different tests or statistical methods, specifying the types of models suitable for biological analysis. and how to validate and evaluate them.

Chapter 5 presents the scientific contributions made in the thesis. It starts with the presentation of the implementation framework for achieving the proposed objectives, continuing with its own approach on the main steps for a complete analysis of gene expressions. Thus, Part I contains the steps for normalizing and reducing the size of the data until the identification of differentially expressed genes, followed by Part II which contains the steps for associating differentially expressed genes in correlation modules, identifying hub genes, and creating graphs.

of co-expression according to the relationships between genes. Part III involves a broad characterization of the genes associated with breast cancer, identified after annotation, followed by Part IV, a modeling of cancer-specific critical genes, in terms of Bayesian networks, to exemplify and identify some possible pathways of cell pathway alteration that can lead to the promotion of breast cancer.

Chapter 6 presents the results obtained from the modeling of the selected gene network as having the highest biological relevance in breast cancer, the critical analysis of the results and the demonstration of the achievement of the objectives. With the help of Bayesian inference, the effects of possible mutations or polymorphisms in each gene, modifying or not the role of the ADAM28 gene in the module, chosen as representative in the tumor progression of breast cancer, were analyzed.

Chapter 7 is a synthesis of the scientific achievements and contributions made, also pointing out the limits of the research undertaken. At the same time, the elements of originality and future development perspectives are indicated. The chapter ends with the presentation of the balance of the research activity carried out during the years of doctoral activity.

The main objective of the doctoral thesis was to identify genes whose expression is correlated with a certain phenotypic trait, in our case breast cancer and which can be used as tools for the realization of possible pathways through which mutations of differentially expressed genes lead at the onset of this neoplasm. The approach described in this paper is useful in discovering and understanding as fully as possible the causal relationships between genes identified as being differentially expressed from a biological data set.

The originality of this paper consists in the integrated analysis of gene expressions from tumor breast tissue. The analysis, performed in the R language, begins with a double filtering of the gene set, with the help of two statistical tests: Chi-square and Welch, at the end of which resulted a small number of differentially expressed genes of statistical significance. Based on these genes, with the help of a correlation analysis, gene modules were identified, from which, based on the correlation link of the top 30 genes in each module, gene co-expression networks were created. From all the resulting networks, the most biologically relevant gene co-expression module (of all the ones we identified) in breast cancer was selected. The schematic reproduction of the chosen module was performed with the help of Bayesian networks, in the Netica development environment. In order to study and identify the main relationships underlying the occurrence of mutations in genes, various probabilities have been assigned to the variables, subjecting the model to various influences, in order to determine the characteristics that the system may have before developing cancer. The emphasis was mainly on modeling the role of the ADAM28 effect gene in the network, in order to conclude on the possible reasons for the occurrence and promotion of breast cancer, being known that mutations in the ADAM28 gene induce cancer progression through various mechanisms.

Chapter 2. Gene expression

2.1. Bioinformatics and data mining

Bioinformatics is a very active and attractive field of research, with a high impact in the development of new technologies, which combines various fields such as biology, medicine, mathematics or computer science [1, 2]. It is defined as research, development and application of computational tools for the use of biological, medical, behavioral or health data [2]. From the perspective of information technology, bioinformatics is a scientific discipline that includes the acquisition, storage, processing, analysis, interpretation and visualization of biological information [3].

2.2. Regulating gene expression

The genetic material is contained in a structure called a chromosome. Each chromosome consists of a long string of DNA associated with proteins, representing a fused chromatin, ie a complex of DNA and proteins. Their most important function is to contain genes. Specifically, for humans, there are 23 pairs of DNA molecules that form chromosomes [4]. Each human cell has 46 chromosomes, which are grouped in pairs, resulting in 23 pairs, of which, from each pair, one chromosome comes from the mother and the other from the father. Each chromosome contains a centromere and two telomeres, each containing a "spiral", a single continuous strand of DNA, whose role is to store and properly transmit hereditary information during cell division.

The gene is the segment of DNA that has the instructions to form a protein. A gene is a sequence of nucleotides in a DNA molecule used to make proteins and RNA. The DNA molecule has the shape of a double helix, the double-stranded structure of DNA (double helix) being ensured by the hydrogen bonds that are established between the nitrogenous bases of the two strands of DNA [5]. The nucleotides in one row bind to the others in the second row as follows: adenine binds to thymine and vice versa, and guanine binds to cytosine and vice versa. The two DNA strands are complementary and antiparallel, their reading being done in the 5 '- 3' direction, from the phosphate end to the hydroxyl. During division, the DNA breaks in half, and each of the two strands of DNA is used as a template to form the complementary part in DNA duplication. The "steps" of the original DNA molecule, in the form of a double spiral, break in place of the hydrogen bond and thus appear two halves of "step" whose nucleotides join with the free ones, restoring the steps of the molecule, resulting in two new molecules and identical to the original DNA, because the sequence of the bases is the same.

Nitrogen bases are molecules that make up nucleotides, RNA, and DNA. DNA and RNA molecules contain five nitrogenous bases: adenine (A), cytosine (C), thymine (T), guanine (G), and uracil (U). The first four nitrogenous bases (A, C, T and G) are part of DNA, while in the composition of RNA, thymine (T) is replaced by uracil (U).

Understanding the structure of DNA is the first step in understanding how biological information is contained in genes.

The process of regulating gene expression is exemplified in Figure 2.1 and captures the framework of protein synthesis, in which two processes take place, one transcription, representing the synthesis of RNA molecules and one translation, involving only the coding DNA, ie genes in which an RNA molecule directly synthesizes a protein.

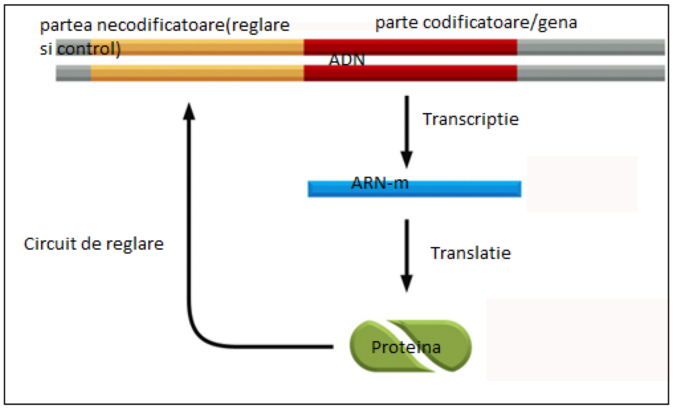


Figure 2.1. Protein biosynthesis [5]

During the nucleus transcription process, the genetic information in the DNA is copied into the messenger RNA in the 5'-3 'direction, with the modification of the thymine nitrogen base, which will be replaced in the RNA by uracil (T becomes U). In other words, transcription occurs when the coding portion of a gene is "rewritten" into a complementary RNA strand called a messenger RNA (mRNA) and is made by a protein complex called RNA polymerase that binds the promoter region of the gene and then "walks" along DNA, catalyzing the formation of mRNA from nucleotide precursors. Transcription factors control the regulation of genes by binding to specific portions of DNA. Binding of the transcription factor can lead to the activation or suppression of transcription, either by causing structural changes in DNA or by interactions with proteins that directly transcribe DNA into mRNA. Binding site transcription factors are relatively short nucleotide sequences (usually 5-15 nucleotides in length) [6].

At the level of the translation that takes place at the level of the ribosomes, the decoding of the messenger RNA takes place, this level being responsible for the protein synthesis. The ribosome, a very complex structure, composed of 2/3 RNA and 1/3 proteins, reads the nucleotide sequence and produces a polypeptide chain, so that in the ribosomes will reach groups of three nucleotides called codons, each codon corresponding to a certain amino acid (AA).

The two significantly different and complex processes each involve a very large number of biochemical reactions (many of which have not been fully characterized). Alterations in biochemical reactions at these levels can lead to genetic code disorders and thus to the emergence of various diseases, including cancer.

2.3. Gene expression profiling techniques

As knowledge of the biological and biophysical basis of cellular function has increased, opportunities have expanded to advance understanding of the cellular and molecular functioning of organic matter, and to design applications in various fields of medical treatment and diagnosis.

Microarray-based gene expression has propelled our knowledge of molecular biology [7]. First of all, microarrays with gene expression have become a widely used technique to study the dynamics of biological processes, being miniaturized laboratories for the study of gene expression [6]. Gene matrices measure the level of molecular RNA expression for thousands of genes simultaneously, the technique being a much-needed data collection method for obtaining information related to understanding the complexity of living organisms [8]. We can obtain answers to biological components that interact with each other, providing us with information about a better understanding of the chosen biological system, microarray analysis with various applications in the medical field, from the characterization of benign or malignant tumors, to the evolution or changes diseases or symptoms over time or response to medication and identification of new treatments [9]. The normal cellular transcriptome can be compared to the transcriptome of a specific disease to try to elucidate disease-specific changes. Another application may be the analysis of physiological changes over a lifetime, for example the comparison of a young transcriptome with an old one [10], revealing changes in molecular pathways.

Basically, a DNA microarray is a collection of microscopic dots attached to a solid surface needed to measure gene expression levels. This technology allows researchers to study a large number of genes simultaneously (approximately 21,000 genes in the human genome) [11]. Gene expression matrix data can be analyzed on at least three levels of complexity [9]. The first level is that of individual genes, where it is looked at whether an isolated gene behaves differently in a control situation versus a treatment situation. The second level is that of multiple genes, in which groups of genes are analyzed in terms of common functionalities or interactions. The third level tries to deduce genes and protein networks that are responsible for the observed patterns.

DNA microarray technologies generate many gene expression profiles. Currently, two microarray technologies are used, namely complementary DNA (cDNA) and oligonucleotides

[12]. Both involve hybridization, which differs in the placement of the DNA sequences as well as the length of the sequences.

cDNA techniques have basic lengths of several hundred to several thousand samples. Usually, for most experiments of profiling cDNA microarray gene expression, mRNA from two different sources (such as diseased cells and normal cells) is extracted, purified, and reverse transcribed into the first strand of cDNA sequences. Each batch of cDNA is labeled with a different fluorophore. After labeling, the cDNA samples are mixed and hybridized to the same microarray. After hybridization, a laser microscope is used to scan the spots. Individual dye emissions at each point are recorded and stored. DNA sequences are marked in different colors (the most commonly used being green and red) and located at different points on the surface of the microarray, indicating higher or lower levels of amounts (expressions) of genes in that sample.

Oligonucleotide matrices involve the use of a chip called the Affymetrix GeneChip, where the expression of each gene is measured by comparing the hybridized mRNA sample with a set of samples consisting of 11-20 pairs of oligonucleotides, each 20-30 nucleotides in length. bases). The first type of sample in each pair is called the perfect match (PM) and is taken from the gene sequence. The second type of sample is called a mismatch (MM) and is created by changing the 13th gene in PM to reduce the specific mRNA binding rate for that gene. For each gene (sample set) two intensity vectors are obtained, one for PM, another for MM.

The difference between the two channels is due to the different hybridization: mRNA in oligonucleotide-type microarrays, compared to cDNA in microarrays on which two cell types are hybridized. Also, the cDNA technique is a two-channel technique and the RNA technique a single channel, the types of samples being synthesized directly on the microarray.

Compared to other biology tools, genomic microarrays are platforms that allow easier access to the internal biological mechanisms of cell cultures. However, while large data sets generated by microarrays are a potential goldmine of biological information, their size is what makes data processing a cumbersome task. This can be further complicated by the inevitable batch effects generated when combining different data sets. Moreover, gene expression profiles are dependent on combinations of complex intracellular events, and as such, identifying signals related primarily to the phenotype of interest is a substantial challenge.

Chapter 3. Synthetic presentation of scientific contributions

3.1. Implementation framework

Data from gene expressions can lead to complex applications such as the discovery of new genes, the diagnosis of various diseases, the discovery of drugs or toxicological research.

With such a large amount of data available to the general public, it is essential for a bioinformatics analyst to have the specific knowledge and skills to understand, analyze and interpret this data in the most accurate way possible.

In this paper, we will analyze gene data from oligonucleotide matrices, from chips called Affymetrix GeneChip, in the R programming language, a special medium used for data manipulation, statistical calculation and graphical display, which offers a wide range of statistical techniques (linear, nonlinear modeling, testing, time series analysis, classification, grouping), including a wide collection of data analysis, manipulation and storage tools [13]. In oligonucleotide matrices, the expression of each gene is measured by comparing the hybridized mRNA sample with a set of samples consisting of 11-20 pairs of oligonucleotides, each 20-30 nucleotides in length (base pairs), each gene (sample set) being represented by two intensity vectors, one for PM, another for MM.

Chapter 3 presents the main steps for the biological analysis of gene expressions, including: downloading CEL files, uploading and normalizing data, filtering the data set, finding differentially expressed genes, grouping genes into clusters and analyzing them, annotating gene symbols and finding of biological relevance according to the chosen topic.

Part I contains the microarray analysis steps performed in the R language until the differentially expressed genes are identified, seen as potential tumor markers (reading data expressions, normalization, filtering, size reduction), followed by part II containing the analysis steps and division of differentially expressed genes into correlation modules, identification of hub genes, as well as creation of co-expression graphs according to the relationships between genes. Part III involves a comprehensive analysis of breast cancer-related genes, followed by Part IV, a modeling of cancer-specific critical genes, to exemplify and identify possible pathways to alter cell pathways that can lead to promoting breast cancer.

3.2. Methodology

In this paper, we used the expression data of the GSE48391 file from the NCBI Gene Expression Omnibus database [14] for the analysis of breast cancer gene expression [14], a file from which only cancer expression data from breast tissue were selected. The GSE series contains lists of GSM files of Affymetrix CEL files, which together form a single experiment, in which the GSM files represent sample-level data from the use of a single chip, in the form of PM and MM intensities. In the mentioned GSE series, for each of the samples / patients, there is a set of characteristics with their specific activity values at a certain point in time, at a certain stage of breast cancer, represented in the form of oligonucleotide matrices from a chip called Affymetrix Human Genome U133 Plus 2.0 Array [2HG-U133_Plus_2].

All 81 files containing data on breast cancer were downloaded and analyzed for the purpose of analyzing the main genes that go out of balance and lead to infinite, uncontrolled multiplication, managing to successfully cross the body's barriers.

Before actually starting the data analysis, representations of the data were made before normalization, with raw data, drawing both a boxplot with non-normalized intensity values and the density vs log intensity histogram for non-normalized data, to see after normalization, the effects this technique on gene expression.

Normalization of microarray data is necessary to ensure that the differences in intensity read by the scanner are due to differentiated gene expressions and not due to printing, hybridization or scanning artifacts, understanding the adjustments made for systematic errors introduced by differences in procedures and dye intensity effects etc.

The Robust Multiarray Averaging (RMA) algorithm in the affy package will be used in this work to normalize all CEL files. Consisting of three steps, namely: background correction of PM values on each separate matrix, normalization and summary of gene expression measurement, RMA analysis uses only information from PM samples to estimate distribution parameters and return the estimated signal. The RMA function [15] sets the chips to the same distribution and the same average and calculates the logarithm of the PMs, converting AffyBatch objects to ExpressionSet objects, thus performing a preprocessing of the raw data (a summary of the sample measurements in one measurement per sample). Inside an ExpressionSet array are the gene expression levels of the *i* genes in the mRNA sample *j*, represented by normalized log2 values. After normalization, we can check the effects of RMA through a boxplot with normalized intensity values.

The normalized data size is 54675 (number of genes) x 81 (number of samples), so the issue of high dimensionality is raised and in this case, an important step that needs to be done is to reduce the size. In this analysis, we selected genes that are expressed in at least 5% of the samples, which have a significantly different variance (higher) than the median variance of all sets of samples. Thus, from the expression matrix of the data set, for each gene, we selected only the values of the expression of the genes greater than the 5th decile, which should be expressed in at least 5% of the total number of samples in the matrix. Then, using a Chi-square test and using a chosen threshold (p value <0.01), we applied a qchisq test to choose genes with a higher than average variance. The Chi-square test for variance is a non-parametric procedure in which a distributed Chi-square test statistic is used, used to determine whether a variant of a variable is different from another specified value, ie whether or not there is a link between the resulting value and an expected value.

Using the quantum function qchisq () in R, the current Chi-square distribution was calculated, which will then be compared with the Chi-square statistical test for variance. Thus, it will be decided whether or not the variance of a gene is different from that calculated using a chosen threshold and degrees of freedom. If the value in the statistical test is higher than the value resulting from the application of the qchisq function [16], then the genes will be retained, otherwise not.

The deliverables will be a much smaller number of genes, genes that have the chance to characterize a phenotype that differs from the rest of the population, the new matrix having much smaller dimensions than the original one, following these filtering steps. In this case, the reduction in the size of the data after the two stages, reaches a number of 19847 (genes) in the 81 existing samples.

For the stage of selection of differentially expressed genes, differentially expressed genes will be identified and selected, with a statistically significant p value, in order to determine if there are subgroups of patients with different profiles (different molecular markers).

Subsequently, the dendogram resulting from the hierarchical grouping using the hclust function was cut into a number of three clusters, based on which they identified the genes differentially expressed between a given cluster and all the others, using a Welch type t test, designed for uneven variations. of samples and / or different sample sizes, but with the presumption of normal distribution. In other words, we tested whether the average of the first cluster is equal to or different from the averages of the other clusters, in which case a “single-tail” test is applied, an alternative way of calculating statistical significance, used for asymmetric distributions, which also calculates value the level of statistical significance p [17]. The lower the p-value, the stronger the evidence that the null hypothesis should be rejected, so that a p-value less than 0.05 is statistically significant. Based on this criterion, we selected the list of genes expressed significantly differentially, establishing a threshold value <0.05, resulting in a number of 7189 genes x 81 samples.

Next, based on the newly resulting matrix, we used the R-language Weighted Correlation Network Analysis (WGCNA) package [18], which forms groups of genes correlated with each other, resulting in relational modules in which the “leading” / hub genes are identified and selected. intramodular cells from the consensus modules, defined as the genes with the highest correlation in that module.

Two matrices are constructed, first an adjacency matrix, the standard correlation method being the Pearson correlation to a threshold power chosen according to a selection criterion, with information about the correlation between the values of expressions between genes, followed by the definition of a TOM matrix (Topology Overlap Matrix), of “neighborhood”, which takes into account the topological similarity and similarities between genes, reflected at the level of network topology [19].

Subsequently, a new hierarchical grouping is made for the division of genes that tend to have high connectivity into co-expression modules. Gene modules are groups of genes strongly interconnected in terms of co-expression, their identification can be done with the help of colors, for easier management. Figure 3.1 shows the gene modules correlated in correspondence with the grouped gene map.

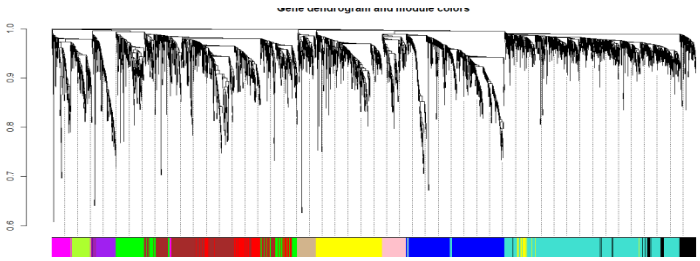


Figure 3.1. Gene grouping according to the TOM matrix and the resulting modules

The minimum number of genes per module was chosen 100, resulting in 13 modules, of which 12 modules with interrelated genes, totaling 3663 genes and a module in which unrelated genes are found depending on distance.

In Figure 3.2, both the co-expressed gene modules and the grouping mode and the relationships between the modules can be seen. Basically, the grouped map of the genes in Figure 3.1 is mapped in the tree in the figure below, of the 12 modules of gene co-expression.

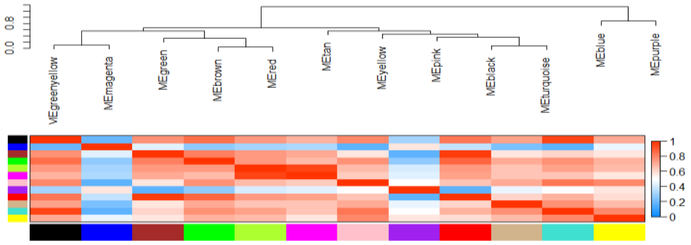


Figure 3.2. Gene co-expression modules grouped according to Euclidean distance

For each module, the adjacency matrices for the genes in question were calculated, and for the verification and validation of the assigned modules, we also opted for the application of the eigengen function, which calculates PCA for data expressions, in which case similar modules are assimilated. The eigengen module is defined as the first major component of a given module, in the form of a one-dimensional data vector, considered a representative of the gene expression profiles / data in that module.

The next step in gene analysis was to identify those genes in each module that are most connected within the module (high intramode connectivity), ie those hub genes most strongly correlated with a clinical or phenotypic trait of interest in our case breast cancer).

In order to find the genes (hubs in each module) with the highest intramodular connectivity, so that the sum of the weights of the edges in the module is maximum, from each adjacency matrix (being an invertible matrix), the maximums were calculated on the column / row for each module separately [20]. Thus, the top 30 genes from each module were chosen, depending on the decreasing values of the results, genes for which new adjacency matrices were created.

Each resulting hub gene in each module was annotated using the "hgu133plus2.db", "AnnotationDbi.db" and "annotated" packages, for each gene id the afferent name / symbol was identified and the Enter id, a unique identifier for each gene, useful for searching for information in genomic databases, such as determining or establishing the biological processes to which they relate.

Using the igraph package in the R programming language, the adjacency matrices for each top 30 genes in each module were read and edge list files were created, representing the correlations of the genes in each module together with their weight. With the help of these files, graphs or gene co-expression networks could be created for each of the 30 correlated genes in each module.

A gene co-expression network (GCN) is an undirected graph in which each node corresponds to a gene and a pair of nodes is connected to an edge if there is a significant co-expression relationship between them.

For each color / module we sorted the genes according to the weight value, selecting the first 30 correlations in the order of decreasing weight values. Entrez_IDs and gene symbols were used to find information about each gene in each module, whether or not it was related to a phenotypic trait of interest, marker, promoter, or gene specific to breast cancer.

Each of the above genes / proteins has been searched for and validated in The Human Protein Atlas [21], as being detected in one or more tissues in the human body, with a greater or lesser degree of occurrence in various types of cancer. From these genes, we selected a number of ten genes with a favorable or unfavorable prognosis for breast cancer [21], namely five genes identified from the gray module, two from the purple module and one from the greenyellow, turquoise and yellow module. For each module, information and features about each gene were extracted.

According to the theory of the hclust function [22], the group with the most correlated genes is on the left side of the dendrogram, in our case the greenyellow module containing the most strongly correlated genes, so the greenyellow module was chosen for further analysis. of the genes and the links between them in the fight against breast cancer. Thus, in addition to RAC2, biological information was also provided for the genes: ADAM28, ARHGAP9, STK17A, CLEC12A, CSF2RB, LY86, TASL, SELL, with an important role in the progression of breast cancer cells in the human body.

Figure 3.3 shows the gene co-expression network (GCN) of the module with the most correlated genes, according to the selection criteria of the co-expression gene analysis package, graph in which each node corresponds to a gene, each pair of nodes being connected with an edge if there is a significant co-expression relationship between the nodes (genes).

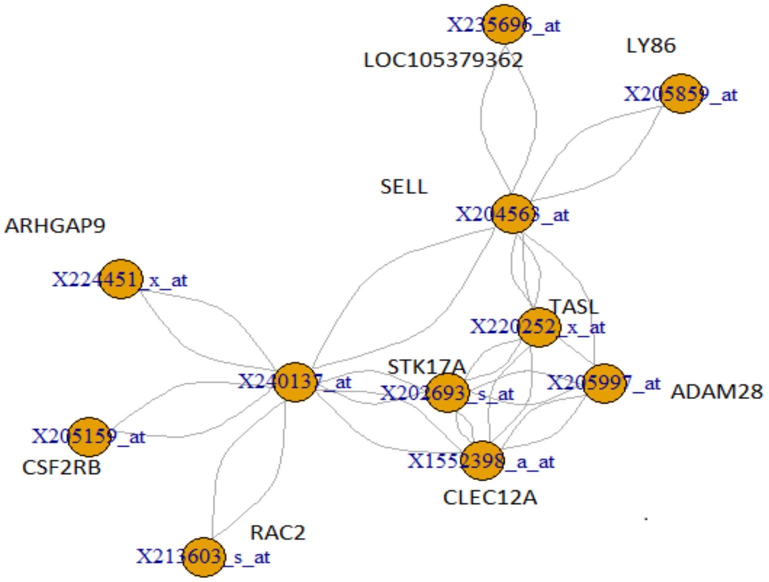


Figure 3.3. Greenyellow module gene co-expression network

As can be seen, the graph is undirected, the meaning of the links between the genes being bidirectional. To exemplify the possible pathways of alteration that lead to the formation of cancer cells, starting from the genes analyzed in Part III, the unidirectional path in Figure 3.3 was chosen.

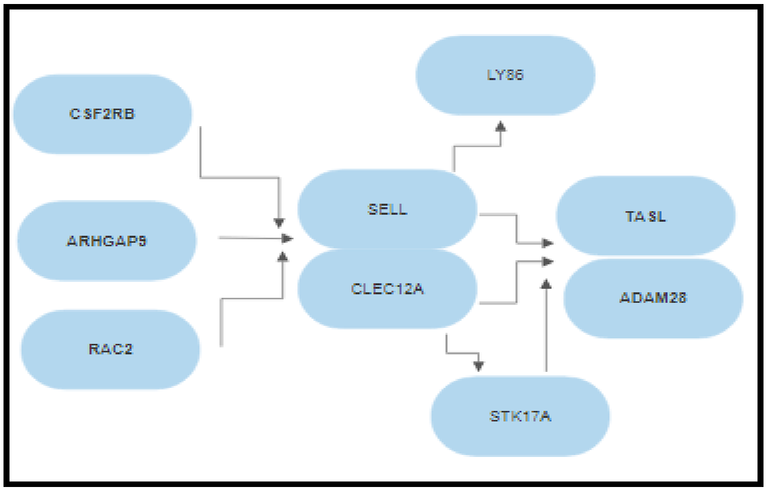


Figure 3.4. Schematic reproduction of the gene system within the module chosen to be modeled

According to the schematic representation above (Figure 3.4), the modeled network starts from the CSF2RB, RAC2 and ARHGAP9 genes to TASL and ADAM28. The main idea of this modeling was to find possible pathways of the mutations of the mentioned genes, in order to give rise to an uncontrolled cell proliferation and later, to an invasion in the neighboring cells.

The main objective of the doctoral thesis was to create pathways through which any mutations that may occur within the genes expressed differently, identified in the stages described above and analyzed in detail with the help of public databases and libraries, to lead to cancer. of the breast. Thus, with the help of Bayesian inference, the effects of possible mutations or polymorphisms in each gene, modifying or not the role of the ADAM28 gene in the module, chosen as representative in the tumor progression of breast cancer, were analyzed.

Bayesian inference is known to be one of the best approaches to uncertainty modeling, with many Bayesian network-based applications being created to model any random variable, including performance indicators in business, engineering, medicine, or ecology. [23] It is based on Bayes' theorem, which can be interpreted as the probability that certain sets of attributes belong to a certain class:

$$P(H_i|S = s_1, \dots s_n) = \frac{P(S=s_1,\dots s_n|H_i)P(H_i)}{P(S=s_1,\dots s_n)} = \frac{P(S=s_1,\dots s_n|H_i)P(H_i)}{\sum_j P(H_j)P(S=s_1,\dots s_n|H_j)}, \qquad (3.1)$$

where $H = \{H_1, \dots, H_n\}$ represents the classes, and S the attributes to be classified.

Bayesian networks are a class of probabilistic models used to model reasoning under uncertainty. Creating a Bayesian network mainly involves two steps:

- ➔ Construction of the qualitative component, ie the directed acyclic graph, containing directed nodes and arcs (which implies the analysis of the problem and the deduction of the causal relations between the variables);
- ➔ Defining the quantitative component, assuming the tables of conditional probabilities attached to each node of the network and describing the uncertainty on the (in) dependence relations between the respective node and its direct parents.

Once the conditional probabilities of the variables have been specified for all possible parent combinations, the Bayesian networks have the property that they will uniquely define a composite of the probability distribution composed over the set of domain variables:

$$P(X_1, \dots, X_n) = P(X_1)P(X_2|parent\_2) \dots P(X_n|parent\_n) = \prod_{i=1}^{n} P(X_i|X_{PA_i}) \qquad (3.\ 2)$$

Netica is an important Bayesian network development software, designed to be powerful and easy to use, an analysis tool chosen by many of the world's top companies and government agencies. [24] Software is a simulation environment for building and evaluating Bayesian networks, with the advantage of an intuitive interface that provides computational support for achieving many types of inference [25, 26]. With the help of Netica, a targeted gene system was created, consisting of nodes and arcs, in which the nodes represent the genes, and the arches the dependency links between the genes.

In our case, the Bayesian network was created exclusively from genes susceptible to breast cancer, identified by R analysis and validated by databases containing information about human cancers [27]. Compared to the diagram in Figure 3.4, the SELL gene with the related offspring was eliminated, the rest of the genes being actively involved (together or separately) in breast tissue neoplasms, according to research and studies conducted so far (Figure 3.5).

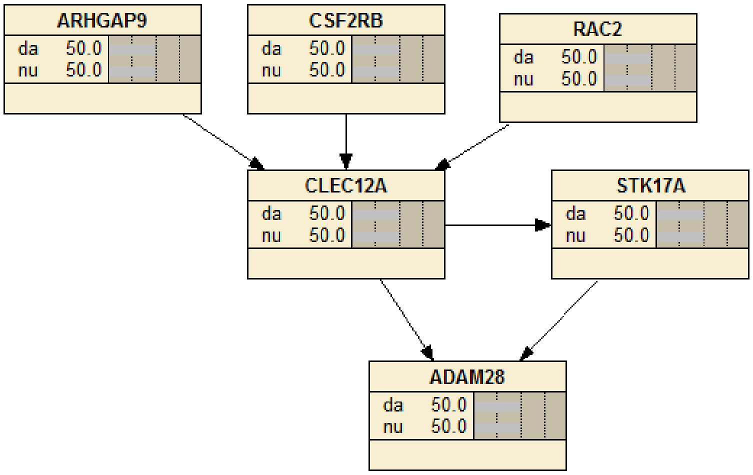


Figure 3.5. The Bayesian network chosen to be analyzed

The qualitative component of the Bayesian network includes all six nodes and the (in) dependence relations between them, meeting in this graph all three types of connections: serial, convergent and divergent. Each gene is, in fact, a node in the Bayesian network created, so to model the network we need the quantitative component, ie both the marginal probabilities of the variables attached to the nodes and the conditional probability tables associated with the nodes, according to the Markov causal condition [28].

In the first phase, we assumed each gene to have equal marginal probabilities, after which we modeled the network for different probabilities to analyze which are the main mutations in genes that can lead to transformations in the structure of the ADAM28 gene. ADAM28, referring to all the other genes in the process.

For our case, the composite probability distribution has the form:

$$\begin{aligned} &P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28) \\ &= P(ARHGAP9)P(CSF2RB)P(RAC2)P(CLEC12A|ARHGAP9, CSF2RB, RAC2) \\ &\qquad P(STK17A|CLEC12A)P(ADAM28|CLEC12A, STK17\mathrm{A}) \end{aligned} \tag{3. 3}$$

and the probability of mutations in the ADAM28 gene, given the other genes, can be written mathematically using the Markov causality condition, resulting in the ADAM28 and RAC2, CSF2RB, ARHGAP9 genes being conditionally independent:

$$P(ADAM28|ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A)$$
$$= \frac{P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)}{\sum_{ADAM28} P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)}$$
$$= P(ADAM28 \mid CLEC12A, STK17A)$$

(3.4)

Next, the influences of certain and uncertain records on the different nodes for all types of connections encountered in the Bayesian network will be presented and analyzed.

Thus, we assumed in turn certain information on mutations in the CLEC12A gene, certain information on mutations in the STK17A gene, followed by cases with definite information on mutations in both genes, to see what are the relationships between genes and what (in) dependence relationships conditional exists between the ADAM28 effect gene and the cause nodes.

The first assumption was that of certain information in the CLEC12A gene, so the probability that the ADAM28 gene is mutant, knowing the mutant CLEC12A gene is:

$$P(ADAM28|CLEC12A = yes)$$
$$= \sum_{STK17A} P(STK17A|CLEC12A = yes) P(ADAM28 \mid STK17A, CLEC12A$$
$$= yes)$$

(3.5)

From the compilation of the multi-connected network (Figure 3.6), we could see that:

- ➔ the modified marginal probability at the evidence function over the probability distribution (100 0) of the CLEC12A gene, leads to changes in the probabilities for the "parent" genes RAC2, CSF2RB, ARHGAP9, but also for the "child" gene STK17A;
- ➔ the marginal probability of the ADAM28 gene changes to the evidence function over the probability distribution (100 0) of the CLEC12A gene;
- ➔ any reliable information on the "parent" genes RAC2, CSF2RB, ARHGAP9, knowing the evidence function over the probability distribution (100 0) of the CLEC12A gene,

does not change the marginal probability of the ADAM28 gene, so we can state that giving the information related to the CLEC12A gene is argued, the initial genes and the final gene are conditionally independent.

➔ any definite information on the "parent" genes RAC2, CSF2RB, ARHGAP9, giving certain information on the CLEC12 gene, does not change the marginal probability of the STK17A gene (the initial genes and the STK17A gene are conditionally independent).

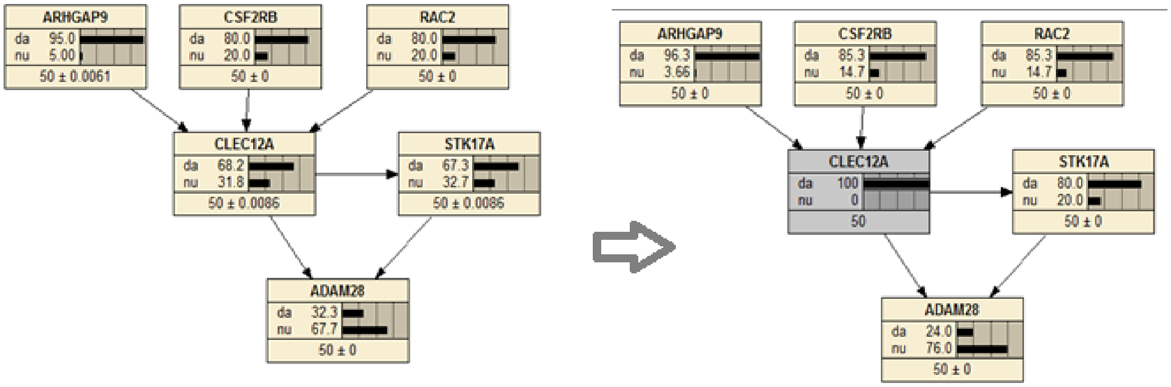


Figure 3.6. Network modeling for definite information in the CLEC12A gene

The second assumption was that of certain information in the STK17A gene, so the probability that the ADAM28 gene is mutant, knowing the mutant STK17A gene is given by the formula:

$$P(ADAM28|STK17A = yes\,) = \sum\nolimits_{CLEC12A} \begin{matrix} P(CLEC12A|ARHGAP9, CSF2RB, RAC2\,) \\ P(ADAM28\,\big|\,CLEC12A, STK17A = yes) \end{matrix} \tag{3.6}$$

From the compilation of the network (Figure 3.7), we could see that:

➔ the modified marginal probability at the evidence function over the probability distribution (100 0) of the STK17A gene, leads to changes in the probability of the CLEC12A "parent" gene, but also in the marginal probabilities of the ARHGAP9, CSF2RB, RAC2 genes;

- ➔ the marginal probability of the ADAM28 gene changes to the evidence function over the probability distribution (100 0) of the STK17A gene;
- ➔ any certain information brought about the “parent” genes RAC2, CSF2RB, ARHGAP, knowing the evidence function over the probability distribution (100 0) of the STK17A gene, modifies the marginal probability of the CLEC12A gene, and therefore of the ADAM28 gene;

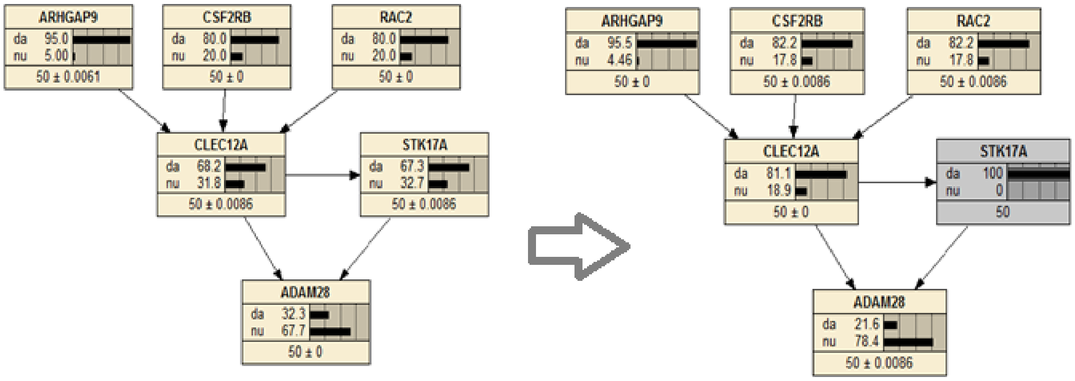
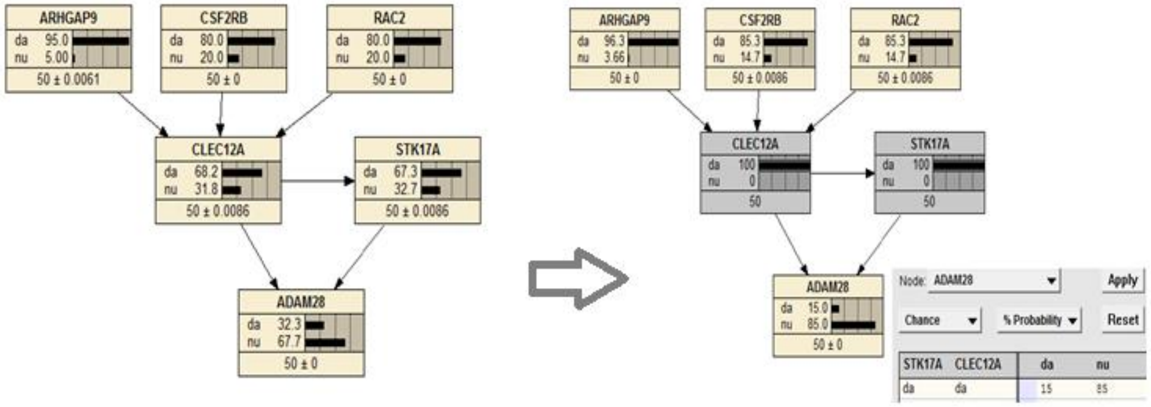


Figure 3.7. Network modeling for definite information in the STK17A gene

The third assumption was that of certain information in both STK17A and CLEC12A genes, so the probability that the ADAM28 gene is mutant, knowing that the two genes are mutant, is given by the data in the conditional probability table for the ADAM28 gene. in which $STK17A = yes, CLEC12A = yes$.

From the compilation of the network (Figure 5.8), we could see that:

- ➔ the marginal probabilities changed at the evidence function over the probability distribution (100 0) of the STK17A and CLEC12A genes, leads to changes in the marginal probabilities of the ARHGAP9, CSF2RB, RAC2 genes (changes appear immediately after certain information of CLEC12A gene, and for the STK17A gene without changing the resulting probabilities once again;
- ➔ the marginal probability of the ADAM28 gene changes to the evidence function over the probability distribution (100 0) of the STK17A and CLEC12A genes and depends only on the line in the conditional probability table for the ADAM28 gene [$P(ADAM28 \,|\, STK17\text{A} = yes, CLEC12A = \text{yes})$].



STK17A	CLEC12A	da	nu
da	da	15	85

Figure 3.8. Network modeling for certain information in the STK17A and CLEC12A genes

Chapter 4. Verification and validation elements. Critical analysis of results

The originality of this paper consists in modeling a network of gene co-expression, with biological relevance in breast cancer, the gene mode chosen after an integrated analysis in the language of R a gene expressions in the GSE48391 file, based on a different approach than the existing ones. For the detection and validation of molecular biomarkers for breast cancer, we reduced the number of genes by double filtering until the identification of those genes differentially expressed with a threshold value that attests to a statistical significance. The WGCNA package was used to identify the key modules and genes representative of our study. Based on correlations and topological similarities between genes, gene co-expression networks have been constructed in order to identify biomarkers (gene hubs) associated with the progression of breast cancer.

In this paper, from the standard expression matrix of data set GSE48391, we selected genes expressed in at least 5% of samples with signal strength greater than the 5th decile of gene expression values, with a variance greater than the median , using a threshold p <0.01 (Chi-square test). Then, using a Welch-type t-test, based on a grouping of samples in a number of three clusters, using the Euclidean distance, the genes expressed differentially were determined based on a p value <0.05. In our case, the reduction in the size of the data after the two stages reaches from a number of 54675 genes, to 19847 genes at the end of the first filtration and to

7189 genes at the end of the second. The 7189 genes were divided into 13 correlation modules, using WGCNA logic, of which 12 modules with correlated genes and a module in which uncorrelated genes are found depending on distance.

The first step in verifying and validating the assigned modules was done by applying the eigengen function, which calculated the PCA for the data expressions. The eigengen module is defined as the first major component of a given module, in the form of a one-dimensional data vector, and considered a representative of the gene expression profiles / data in that module. Figure 4.1 shows the validation of the 12 modules identified in Chapter 3, the difference between the dynamic grouping and the eigengen function being the assimilation of the similar modules of the latter (for example, the magenta module is assimilated by the greenyellow module, the red one by the brown module). turquoise assimilated by the black module).

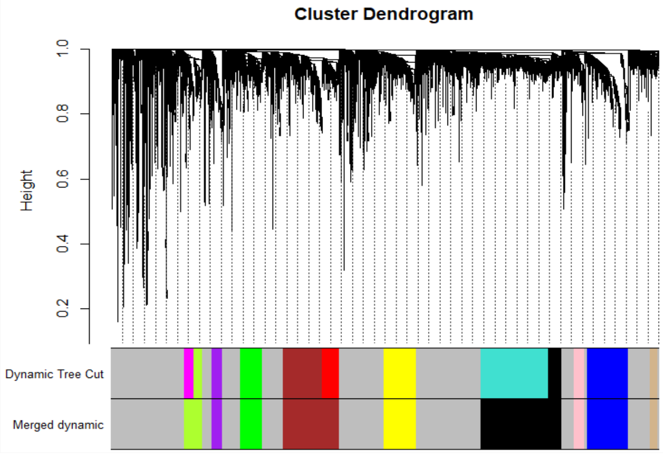


Figure 4.1. Validation of assigned modules using the eigengen function

Other independent data sets were used to validate the genes in our analysis (GSE48391). To do this, we downloaded the GSE36295 file [29], which contains RNA isolated from surgically excised, purified, labeled, and hybridized breast cancerous tissue on the Affymetrix Human Gene 1.0 ST Array platform. For the GSE36295 file, the number of genes decreased from 32321 x 50 to 8594 x 50 and then to 3685 x 50. From 32321 genes and 50 samples, after filtering we reached 3685 genes, from which a number of genes were selected from the module with the genes those more related: SKAP2, ITGB5, MYCT1, PODNL1, TEK, SDR42E1, TFPI, JPH1, PAPSS2, MANEAL, PPIC, COBLL1, EMCN, all tumor markers in various cancers, but the gene of interest ADAM28 could not be found in the matrix normalized gene expressions.

Another GSE102907 file [30] was downloaded, containing the messenger RNA extracted from the primary tumor of some breast cancer patients, hybridized and scanned with the Affymetrix Human Genome Matrix GeneChip U133 Plus 2.0. From 54675 genes and 61 samples, after filtration we reached 4541 genes, from which a number of genes were selected from the module containing the most correlated genes: MIPOL1, CDC20B, C7orf57, MPV17L, ZBTB18, CYP4F8, MYBPC1, KITLG , FAM110C, CSTF3, CROT, ARMC3, the ADAM28 gene being found as a differentially expressed gene.

Biological knowledge from databases, such as the Atlas of Genetics and Cytogenetics in Oncology and Haematology [31], which contains information on all gene annotations, and the Human Protein Atlas [21], have been used for a better understanding of the structure. , the location and characteristics of those genes, all of which are susceptible to mutations in breast cancer. In addition, the clinical information on breast cancer in the Catalog of Somatic Mutations in Cancer (COSMIC) [32] was used to validate genes identified and analyzed by us as related to the phenotype of interest in the module chosen to be modeled, and different Databases such as GEO, COSMIC, Uniprot or Protein Atlas have been used to analyze the characteristics of each gene and how or not they relate to the proliferation of altered cells in the body.

During the study, a survival analysis of the hub genes identified in Chapter 3 was used using the Kaplan Meier plotter. All genes analyzed in the chosen gene network were identified in the Kaplan Meier plotter application as biological biomarkers, one of the main purposes of this application being to validate the biomarkers identified by various techniques.

Many papers state that the ADAM28 gene is overexpressed in several cancers, including breast cancer [33, 34], but none of the microarray analyzes treated in publicly available works have identified the ADAM28 gene as a possible target gene in breast cancer. Instead, our analysis, in addition to selecting as a primary cause of the chosen module, a prognostic marker [35] (whose presence and change in concentration is correlated with the development of tumors), namely the RAC2 gene, also identified the gene effect as a communicator with the body's immune cells - ADAM28 and possible ways of transmitting mutations between several genes involved in the processes of cell growth and proliferation in the human body. Thus, the main objective of the doctoral thesis, that of making possible routes through which any mutations that may occur in the genes, to lead to breast cancer, has been achieved.

The emphasis was mainly on modeling the role of the ADAM28 effect gene in the network, in order to conclude on the possible reasons for the appearance and promotion of breast cancer, being known that mutations in ADAM28 induce cancer progression through various mechanisms. On the other hand, the gene is considered a potential therapeutic target, being known to normally play a protective role against the spread of cancer cells by promoting T cells [36, 37], so that a better understanding of network interactions in which is part of it is vital to establish proper procedures for the administration of possible treatments to ensure the proper functioning of the gene in the body.

Chapter 5. Synthesis of scientific contributions and prospects for future development

5.1. Synthesis of scientific contributions

Breast cancer is one of the leading causes of death for women worldwide. The aim of the paper was to create and model gene networks that have biological relevance in this type of cancer. The interpretation of gene expression values consisted of a first phase in a microarray analysis in the R language, followed by a second phase, that of the analysis and validation of identified genes and the search for biological information about them in public space, as finally the network. gene containing the most distantly correlated genes, to be analyzed and modeled in order to find and interpret possible pathways for mutations in breast cancer-specific genes (in addition to those known widely: ex BRCA1, BRCA2) - Figure 5.1.

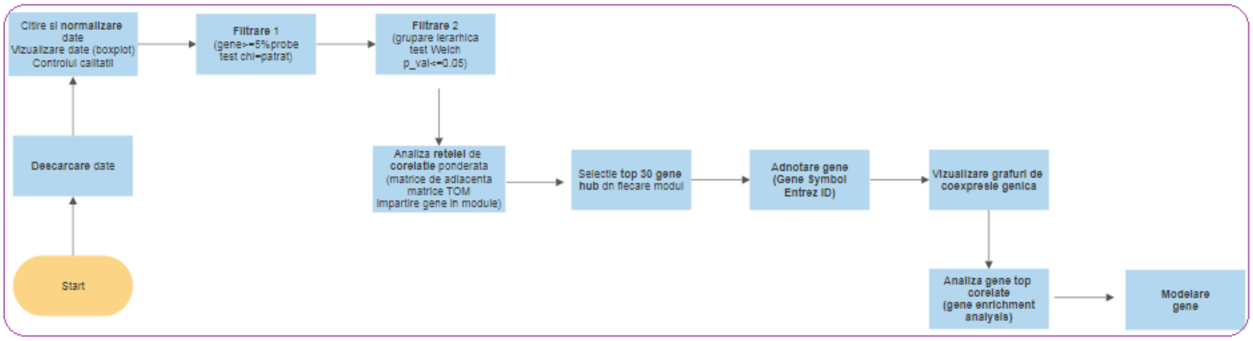

Citire si normalizare date
Vizualizare date (boxplot)
Controlul calitatii
Filtrare 1 (gene>=5%probe test chi=patrat)
Filtrare 2 (grupare ierarhica test Welch p_val<=0.05)
Descarcare date
Start
Analiza retelei de corelatie ponderata (matrice de adiacenta matrice TOM impartire gene in module)
Selectie top 30 gene hub dn fiecare modul
Adnotare gene (Gene Symbol Entrez ID)
Vizualizare grafuri de coexpresie genica
Analiza gene top corelate (gene enrichment analysis)
Modelare gene

Figure 5.1. Proposed microarray analysis steps

In short, the steps followed were:

➔ installation of packages required for microarray analysis in R language: "BiocManager", "affy", "GEOquery", "affyPLM", "hgu133plus2.db", "AnnotationDbi.db", "annotate", "dynamicTreeCut", "WGCNA", "igraph";

➔ downloading CEL files from the NCBI Gene Expression Omnibus database (GSE48391 - oligonucleotide matrices from the Affymetrix Human Genome U133 Plus 2.0 Array);

➔ loading and reading data in the R programming language;

The raw CEL files of each microarray dataset were imported into R (version 4.1.0) using the ReadAffy function of the BiocManager 3.13 affy package.

➔ normalization of data with the RMA method, which involves three steps: correction of background noise, normalization of distributions and transformation of expressions into logarithms in base 2, only based on PM values in matrices;

We chose a different approach, related to filtering the number of genes, than the usual approaches. My approach consists of two parts:

➔ filtering no. 1 of the sample set by selecting genes expressed in at least 5% of the samples, with a significantly different variance from the median (Chi-square test);

➔ filtering no. 2 of the sample set: identification of differentially expressed genes based on p values, using a statistical test (Welch t-test);

➔ analysis of the weighted correlation network, by identifying co-expressed gene modules using the WGCNA package;

➔ annotating gene sets to genetic symbols (Enter Gene ID and Symbol);

Annotations and information from the public biological space showed that genes of biological relevance were found in all of the 13 modules in several types of cancer, and in 5 (gray, greenyellow, yellow, purple, turquoise) of the 13 modules were identify markers that predict breast cancer.

➔ identification of the most connected genes in each module (genes with high intramode connectivity), ie those hub genes most strongly correlated with the chosen process (in our case breast cancer);

➔ making gene graphs / gene co-expression networks using the igraph package (gene regulatory networks) for the most correlated 30 genes in each module;

- ➔ identifying the characteristics of genes with a role in the development of breast cancer from all the resulting gene modules;
- ➔ analysis and characterization of a single gene module based on hierarchical grouping (and moreover with the highest biological relevance for our study), the module containing genes with similar expression profiles and important functions, with biological relevance in the progression of breast cancer;

With the help of several public databases we extracted biological knowledge about genes and we checked the biological relevance of the genes in the chosen module, performing a biological analysis (gene enrichment analysis) on the obtained network, to see if the module has biological meaning or not. The information found was brief:

- o ARHGAP9: GTPase activator plays essential roles in regulating cell growth, cell differentiation, cell migration, related to Cdc42, a protein involved in regulating the cell cycle and RAC1, which regulates several signaling pathways that control cell organization, transcription and proliferation;
- o CSF2RB is a receptor for GM-CSF, a growth factor that induces differentiation and proliferation in the bone marrow;
- o RAC2 has a role in regulating cellular responses, such as apoptotic and epithelial cell processes, being involved in mutations;
- o CLEC12A encodes proteins with a role in cell signaling and the immune response (some of the proteins being located in the region of the "killer gene" on chromosome 12p13);
- o STK17A is a member of cellular apoptosis;
- o ADAM28 is known to be linked to immune system cells, with a protective role by promoting T cells in the body;

- ➔ visualization of proteins encoded by selected genes and specific regions, such as the location of mutations that may occur within them;
- ➔ modeling the gene module / co-expression network using the Bayesian probabilistic inference, in the form of a directed acyclic graph (Bayesian network);
- ➔ identification of possible pathways / pathways for transmitting mutations to the ADAM28 gene, a gene with biological significance for binding to the body's immune cells;

- ➔ identification of rules for mutations in the ADAM28 gene, depending on the type of connection in the Bayesian network and the type of information provided (or not) on a particular gene in the system;

The main objective of the doctoral thesis was to identify genes whose expression is correlated with a certain phenotypic trait, in our case breast cancer and which can be used as tools for the realization of possible pathways through which mutations of differentially expressed genes lead at the onset of this neoplasm. Thus, Bayesian networks were chosen for modeling the gene co-expression network, which were useful in modeling reasoning under uncertainty. In addition, they have an intuitive and flexible language for representing the dependencies and independences between the variables of the chosen module. Both components of the Bayesian network, quantitative and qualitative, are transparent, in the sense of having complete information on probability values and continuous observation of dependencies between nodes, making the gene scheme very suggestive. Valuable information can be extracted through it and the values applied, for the purpose of analyzing (inter) dependencies between nodes. In addition to these characteristics, Bayesian networks allow the introduction of several types of reasoning, compared to other types of systems that allow only one.

Depending on the existence of a certain type of evidence or information (diagnostic tests, ultrasounds, blood sampling, etc.) on the genes, we can provide additional information about their changes in the chosen biological process. Thus, if there is no definite information from the physician or expert, Bayesian network modeling is performed relative to previously known a priori probabilities (usually calculated using the total probability formula), resulting in new marginal a priori probabilities. If this certain information is known, the a posteriori probabilities of the nodes are calculated. In our case, the influence of certain information on each node in the system was analyzed, resulting in a number of seven cases of probabilistic inference of causal or predictive reasoning:

a. If the information related to the ARHGAP9 "parent" gene is true, then:

- ➔ the probabilities of the other two "parent" genes do not change;
- ➔ the probability of the CLEC12A gene changes;
- ➔ the change suffered by the CLEC12A gene leads to changes in the STK17A and ADAM28 genes;
- ➔ the probabilities of the other two "parent" genes do not change;

b. If the information related to the CSF2RB "parent" gene is true, then:

➔ the probabilities of the other two "parent" genes do not change;

➔ the probability of the CLEC12A gene changes;

➔ the change suffered by the CLEC12A gene leads to changes in the STK17A and ADAM28 genes;

c. If the information about the RAC2 "parent" gene is true, then:

➔ the probabilities of the other two "parent" genes do not change;

➔ the probability of the CLEC12A gene changes;

➔ the change suffered by the CLEC12A gene leads to changes in the STK17A and ADAM28 genes;

d. If the information related to two or all of the RAC2, CSF2RB, ARHGAP9 "parent" genes is true, then:

➔ the probability of the CLEC12A gene changes and increases significantly compared to the cases a., b., c. (P(CLEC12A | RAC2=yes, CSF2RB=yes, ARHGAP9 =yes));

➔ the change suffered by the CLEC12A gene leads to changes in the STK17A (probabilities decrease compared to the initial values) and ADAM28 (probabilities increase compared to the initial values) genes;

e. If the information about the CLEC12A gene is true, then:

➔ the probabilities of the "parent" genes increase compared to the initial values;

➔ STK17A gene changes depending on (P(STK17A| CLEC12A=yes));

➔ ADAM28 gene changes depending on (P(ADAM28| STK17A, CLEC12A=yes));

➔ any (certain) information on the “parent” genes RAC2, CSF2RB, ARHGAP9 does not change the confidence in the CLEC12A gene;

➔ any information on the “parent” genes RAC2, CSF2RB, ARHGAP9 does not change the marginal probability of the STK17A gene (the initial genes and the STK17A gene are conditionally independent).

➔ any information on the “parent” genes RAC2, CSF2RB, ARHGAP9 does not change the confidence (marginal probability) in the ADAM28 gene, so we can say that the initial genes and the final gene are conditionally independent.

f. If the information about the STK17A gene is true, then:

➔ the probabilities of the initial genes decrease compared to the initial values;

➔ the probability of the CLEC12A gene decreases compared to the initial value P(CLEC12A);

➔ ADAM28 gene changes depending on (P(ADAM28| STK17A=yes, CLEC12A));

➔ any (certain) information provided on one or more “parent” genes (RAC2, CSF2RB, ARHGAP9) changes the confidence in the CLEC12A gene, and the probability increases compared to the initial one;

➔ the change suffered by the CLEC12A gene also leads to changes in the ADAM28 gene (increased probabilities);

g. If the information related to both CLEC12A and STK17A genes is true, then:

➔ ADAM28 gene changes depending on (P(ADAM28| STK17A=yes, CLEC12A=yes));

➔ any (certain) information on the “parent” genes RAC2, CSF2RB, ARHGAP9 does not change the confidence in the CLEC12A and STK17A genes;

➔ any (certain) information provided on the RAC2, CSF2RB, ARHGAP9 “parent” genes does not change the confidence in the ADAM28 gene, so we can state that the initial genes and the final gene are conditionally independent.

Building intuitive graphic designs has become a popular technique that allows us to better understand different processes, in our case a possible biological pathway involved in breast cancer networks. One of the most important advantages of these graphs is the versatility of the models, with the mention of respecting the reliability of the system.

All of the above cases belong to predictive or causal reasoning, in which we assume the known causes and want to see what their influence is on the effect gene (what is the cause X that produced the Y effect). The notion of cause must be interpreted in the sense of a factor which may lead to an increase or decrease in the probability of achieving other parameters which it conditions.

The approach described in this paper is useful in discovering and understanding as fully as possible the causal relationships between genes identified as being differentially expressed from the GSE48391 dataset. Noticing the influence of dependency and independence relationships on gene assemblies may be a defining factor in predicting future behaviors of these genes or finding new relationships between them. In this thesis, all genes participating in the analyzed Bayesian network were validated by biological databases as having a certain link to cell proliferation, immune cells or other pathways that could be the target of dangerous cellular alterations on the proper functioning of the human body. The model created in the Netica development environment respected the gene co-expression network offered by the igraph package and chosen by us as having the highest biological relevance for the study, keeping a single sense of gene linkage, starting with RAC2, CSF2RB and ARHGAP9 genes and ending with the ADAM28 gene, known as a protease whose mutations induce cancer progression through various mechanisms of cancer cell proliferation. The analysis in Netica involved multiple probabilistic inferences on the connections between the nodes in the network, in order to interpret the possible pathways of mutation succession, starting from the baseline (cause genes) and continuing to lower levels (effect gene involved in tumor growth and metastases).

Thus, I was able to conclude on the following points:

- ➔ If we do not have (certain) information on the RAC2, CSF2RB, ARHGAP9 genes, all the links in the system must be taken into account (this step is the most complicated, because it involves every node and link in the system).
- ➔ If we do not have any (certain) information on the CLEC12A and / or STK17A genes, the link between the initial genes in the network and the final gene must be taken into account. the structure of the neoplastic breast tissue of these genes, to try to stop or perpetuate subsequent mutations in the network or at least reduce the rate of mutations).
- ➔ If we have definite information on one or both of the CLEC12A and STK17A genes, then the link between the initial network genes and the final gene is no longer of interest (being conditionally independent), and we can intervene biologically, by silent and bounded expression. normal mutations, which no longer allow the invasion of healthy cells by cancer cells) only on the parents of ADAM28.
- ➔ If we have (certain) information on the RAC2, CSF2RB, ARHGAP9 genes, then the links between the three key genes in the network must be rigorously analyzed: CLEC12A, STK17A and ADAM28.
- ➔ If we do not have (certain) information about the CLEC12A and / or STK17A genes, any change in one of the parent genes (CLEC12A and STK17A) leads to changes in the child gene (ADAM28). In this case, both CLEC12A and STK17A genes should be analyzed to identify critical areas where mutations that promote proliferation to new genes occur.
- ➔ If we have definite information on one or both of the CLEC12A and STK17A genes, then we need to work directly on the ADAM28 gene to try to limit the increase in gene

mutations and the proliferation of altered cells in other important DNA segments involved in tumor growth. our case, breast cancer.

As a limitation of the analysis and identification of possible pathways of mutations in genes susceptible to control and regulation of cell cycle progression and apoptosis, we can specify that the probabilities of nodes were chosen randomly, real data about these genes having a much stronger and more conclusive impact on those determined by us, but from our searches, they could not be found on public platforms or databases, for a more real analysis.

Moreover, the availability of papers attesting to the links between the genes analyzed in this thesis could not be identified, our results being validated by identifying those genes whose values are statistically significant, with biological relevance in cell growth, whose mutations lead to multiple transformations in various types of cancer, but especially in breast cancer. The observations and information inside the paper, confirmed by the literature, are separated for each gene and come to help a better understanding of the system. A validation of the gene relationships in the chosen module to be analyzed would be appreciated, but this would be possible only through the complex work in a genetics laboratory, under close supervision and a generalized effort of a mixed team of doctors, engineers, chemists and biologists.

The originality of this paper is an integrated R-language analysis of gene expressions in the GSE48391 file, which is different from the existing ones. The analysis begins with a double filtering of the gene set, with the help of two statistical tests: Chi-square and Welch t-test, at the end of which resulted a small number of differentially expressed genes of statistical significance. Based on these genes, using the WGCNA correlation analysis, thirteen gene modules were identified, of which, based on the correlation link of the top 30 genes in each module, gene co-expression networks were created using the igraph package. From all the resulting networks, the most biologically relevant gene co-expression module (of all the ones we identified) in breast cancer was selected. Multiple databases, such as COSMIC, UniProt, canSAR or Protein Atlas, have been used to analyze the characteristics of each gene and see how they relate to the proliferation of altered cells in the body. The schematic reproduction of the chosen module was performed with the help of Bayesian networks, in the Netica development environment. In order to study and identify the main relationships underlying the occurrence of mutations in genes, various probabilities were randomly assigned to node variables, subjecting the model to various influences, in order to determine the characteristics that the system may have before developing cancer. . The emphasis was mainly on modeling the role of the ADAM28 effect gene in the network, in order to conclude on the possible reasons for the occurrence and promotion of breast cancer, being known that mutations in the ADAM28 gene induce cancer progression through various mechanisms.

5.2. Future perspectives

A first thing we set out to test in the future is another important feature of Bayesian networks, namely the possibility of intervening in the causal process, with situations in which the graph can be increased by one or more intervention variables (new causes). For example, in the case of the current Bayesian network, a new “parent” node of the CLEC12A node can be introduced, in the form of a “treatment” type node and based on it, to analyze its influence on the ADAM28 effect node and the others nodes, how to change trusts in network variables by introducing the new variable "treatment", seen as a factor in improving the effects that mutations can have on the gene-child, resulting in a refresh of genes at lower levels newly emerged genes.

In other words, the comparison of several genes specific to various types of cancer can give us information of great interest about the genes that are found in one or more types of neoplasms, but also those that differ from one type of cancer to another.

Moreover, another future perspective would be to identify the genes differentially expressed from an altered material from a healthy genetic one and compare them in order to identify the differences between the two genetic materials.

Selective Bibliography

[1] N.G. Bourbakis, “Bio-imaging and bio-informatics”, IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 33(5):726–727, 2003.

[2] R. Fuchs, “From sequence to biology: the impact on bioinformatics”, Bioinformatics, 18(4):505–506, 2002.

[3] G.B. Singh, “Fundamentals of Bioinformatics and Computational Biology. Methods and Exercises in Matlab”, Modeling and Optimization in Science and Technologies, Springer, Vol 6, ISBN 978-3-319-11402-6, 2015.

[4] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, "Molecular Cell Biology", 4th edition, New York: W. H. Freeman, ISBN-10: 0-7167-3136-3, 2000.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, "Molecular Biology of the Cell", 4th edition, New York: Garland Science, ISBN-10: 0-8153-3218-1ISBN-10: 0-8153-4072-9, 2002.

[6] J.Ernst, “Computational Methods for Analyzing and Modeling Gene Regulation Dynamics”, School of Computer Science, Machine Learning Department, CMU-ML-08-110, 2008.

[7] J.M. Keith, “Bioinformatics Vol I. Data, Sequence Analysis and Evolution“, Methods in Molecular Biology, Humana Press, ISBN: 978-1-58829-707-5, 2008.

[8] I.P. Androulakis, E. Yang, R.R. Almon, “Analysis of Time-Series Gene Expression Data: Methods, Challenges and Opportunities”, The Annual Review of Biomed.ical Engineering, 9:3.1-3.24, 2007.

[9] P.Baldi, S. Brunak, “Bioinformatics. The machine learning approach”, Second Edition, MIT Press, 2001.

[10] M.R. Speicher, S.E. Antonarakis, A.G. Motulsky, “Vogel and Motulsky’s Human Genetics. Problems and Approaches 4th Edition”, Spinger, 2010.

[11] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, “Hidden Markov models for cancer classification using gene expression profiles”, Information Sciences 316, 293-307, 2015.

[12] J. J. Pasternak, “An introduction to Human Molecular Genetics. Mechanisms of Inherited Diseases”, Second Edition, WILEY-Liss, 2005.

[13] The R Project for Statistical Computing. https://www.r-project.org/.

[14] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/.

[15] K. Taskova, “Introduction to Microarray Analysis”, Computational Biology and Data Mining Group, Faculty of Biology, Gutenberg Universitat. 2016.
https://cbdm.uni-mainz.de/files/2016/02/GE_microarrays.pdf.

[16] Chi-square test for variance.
https://www.empirical-methods.hslu.ch/decisiontree/differences/variance/1-13chi-square-test-for-variance/.

[17] Institute for Digital Research & Education Statistical Consulting.
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/.

[18] P. Langfelder, S. Horvath, "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics, 9: 559, doi: 10.1186/1471-2105-9-559, 2008.

[19] J. Tang, D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong, G. Wu, "Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis", Front. Oncol., https://doi.org/10.3389/fonc.2018.00374, 2018.

[20] S. Horvath, J. Dong, "Geometric Interpretation of Gene Coexpression Network Analysis", PLOS Computational Biology, https://doi.org/10.1371/journal.pcbi.1000117, 2008.

[21] The Human Protein Atlas. https://www.proteinatlas.org/.

[22] Hierarchical Clustering function.
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust.

[23] Top Open Source Tools For Bayesian Networks. https://analyticsindiamag.com/top-8-open-source-tools-for-bayesian-networks/.

[24] Netica software. https://www.norsys.com/.

[25] I. Mihu, "Tehnici de decizie şi diagnoză", Editura Universitară, 2008.

[26] R. Mihu, M. Voinescu, O. Arsene, ”Tehnici de decizie şi diagnoză. Aplicaţii”, Editura Universitară, 2009.

[27] A. Pavlopoulou, D.A. Spandidos, I. Michalopoulos, "Human cancer databases (Review)", Oncol Rep, 33(1): 3–18, doi: 10.3892/or.2014.3579, 2015.

[28] I.-O. Lixandru-Petre, “Tehnici de asistare a deciziei şi diagnozei. Indrumar de laborator”, MatrixRom Bucuresti, ISBN: 978-606-25-0563-9, MatrixRom Bucuresti, 2020.

[29] Expression profiling by array. GSE36295.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36295.

[30] Expression profiling by array. GSE102907. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102907

[31] Atlas of Genetics and Cytogenetics in Oncology and Haematology. http://atlasgeneticsoncology.org/.

[32] Catalogue Of Somatic Mutations In Cancer. https://cancer.sanger.ac.uk/cosmic/.

[33] Y. Mitsui, S. Mochizuki, T. Kodama, M. Shimoda, T. Ohtsuka, T. Shiomi, M. Chijiiwa, T. Ikeda, M. Kitajima, Y. Okada, "ADAM28 is overexpressed in human breast carcinomas: implications for carcinoma cell proliferation through cleavage of insulin-like growth factor binding protein-3", Cancer Res, 66(20):9913-20, DOI: 10.1158/0008-5472.CAN-06-0377, 2006.

[34] C. Gérard, C. Hubeau, O. Carnet, M. Bellefroid, N.E. Sounni, S.Blacher, G. Bendavid, M. Moser, R. Fässler, A. Noel, D. Cataldo, and N. Rocks, "Microenvironment-derived ADAM28 prevents cancer dissemination", Oncotarget; 9(98): 37185–37199, doi: 10.18632/oncotarget.26449, 2018.

[35] Markeri tumorali. https://www.cdt-babes.ro/articole/markeri_tumorali_generalitati.php.

[36] C. Gérard, C. Hubeau, O. Carnet, M. Bellefroid, N.E. Sounni, S. Blacher, G. Bendavid, M. Moser, R. Fässler, A. Noel et al., “Microenvironment-Derived ADAM28 prevents cancer dissemination”, Oncotarget, 9, 37185–37199, 2018.

[37] C. Hubeau, N. Rocks, D. Cataldo, “ADAM28: Another ambivalent protease in cancer”, Cancer Lett., 494, 18–26, 2020.

List of publications

1. Articles published in Web of Science indexed journals

- ➔ I.-O. Lixandru-Petre, C. Buiu, “An integrated breast cancer microarray analysis approach”, U.P.B. Ski. Bull., Series C, Vol. 84, Iss. 2, ISSN 2286-3540, 2022;
- ➔ I.-O. Lixandru-Petre, C. Buiu, “Modeling breast cancer gene expression using bayesian networks”, U.P.B. Ski. Bull., Series C, ISSN 2286-3540, 2022 - accepted for publication;

2. Articles published in Web of Science indexed scientific events

- ➔ I.-O. Petre, C. Buiu, “An integrated gene expression analysis approach”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Iași, DOI: 10.1109 / EHB.2015.7391442, WOS: 000380397900095, 2015;
- ➔ I.-O. Petre, C. Buiu, “A colon cancer microarray analysis technique”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Sinaia, DOI: 10.1109 / EHB.2017.7995412, WOS: 000445457500067, 2017;
- ➔ I.-O. Petre, “Rb protein dynamic modeling”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Sinaia, DOI: 10.1109 / EHB.2017.7995485, WOS: 000445457500140, 2017;
- ➔ I.-O. Lixandru-Petre, “Modeling a Bayesian Network for a Diabetes Case Study”, Proceedings of the International Conference on e-Health and Bioengineering (EHB), Iași, IEEE Xplore, DOI: 10.1109 / EHB50910.2020.9280179, WOS: 000646194100054, 2020;
- ➔ I.-O. Lixandru-Petre, "A Fuzzy System Approach for Diabetes Classification", Proceedings of the International Conference on e-Health and Bioengineering (EHB), Iași, Romania, IEEE Xplore, DOI: 10.1109 / EHB50910.2020.9279882, WOS: 000646194100008, 2020;

3. Articles published in BDI indexed scientific events

- I.-O. Petre, C. Buiu, “Microarray Gene Expression Analysis using R”, International Conference on Advancements of Medicine and Health Care through Technology (Meditech), Cluj-Napoca, IFMBE Proceedings book series (volume 59), DOI: 10.1007 / 978-3 -319-52875-5_74, 2016;

4. Articles presented at international conferences

- I.-O. Petre, “Classifying different subtypes of malignant processes based on gene expression analysis”, The Christie International Cancer Careers Conference, The Christie School of Oncology, Manchester, DOI: 10.13140 / RG.2.2.36691.94242, 2015;