AN INTEGRATED BREAST CANCER MICROARRAY ANALYSIS APPROACH

Irina-Oana LIXANDRU-PETRE[1], Catalin BUIU[2]

As knowledge of the biological and biophysical basis of cellular function has increased, opportunities to understand the cellular and molecular functioning of organic matter have expanded, and gene expression microarrays have become a widely used technique to study the dynamics of biological processes. In this paper, we present a gene expression analysis approach using R programming, starting from mRNA oligonucleotide matrices ending with interrelated genes, analyzed within a chosen biological process.

Keywords: microarray, gene, module, co-expression, analysis

1. Introduction

Cancer is at the top of major noncontagious diseases (including Parkinson's disease, autoimmune diseases, strokes, diabetes, chronic kidney disease, osteoporosis, Alzheimer's disease, or cataracts), responsible for about 15% of all human deaths [1]. Ranging from raw sequence data to well-structured data, transforming and analyzing data for knowledge extraction is a major challenge for researchers.

Various biological databases are used in scientific communities. Among them, we enumerate the National Center for Biotechnology Information database (NCBI) [2] with its public databases of sequences and experiments, Protein Data Bank (PDB) [3], or Universal Protein Resource (UniProt) [4]. The database for nucleic acid research (NAR) [5] provides over 1600 biological databases, including nucleotide, RNA or protein sequences (COG, Pfam, SMART, Panther, PED), metabolic and signalling pathways (STRING, KEGG, KLIFS), gene and genomic databases (Ensembl and UCSC Genome Browser), microarray data or molecular biological databases.

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, starting from the development of methods for deducing the structure and function of genes,

[1] PhD, Dept. of Automatic Control and Systems Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: irina.petre@upb.ro
[2] Prof., Dept. of Automatic Control and Systems Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: catalin.buiu@upb.ro

grouping proteins based on amino acid sequences, ordering and classification of proteins according to the degree of similarity, alignment of sequences, to complex applications such as discovering new genes, better diagnosis of various diseases or discovering up to date drugs.

In this paper, an integrated microarray breast cancer analysis approach is presented. Using Gene expression omnibus (GEO), one of the significant public databases of microarray experiments and RNA profiling, managed by NCBI, GSE48391 file expression data was selected to analyse breast cancer gene expression.

2. Gene expression analysis

A. Introduction to DNA microarrays

The process of analyzing information sequences has propelled our knowledge of molecular biology, being involved in the study of many aspects of biological nature.

Gene regulation is a central and natural biological process, and its disruption can lead to many diseases. The process is controlled mainly by a dynamic network of transcription factors that interact with specific genes to control their expression [6]. A widely used technique for studying the dynamics of biological processes and gene expression is gene microarrays. DNA microarray is a collection of microscopic dots attached to a solid surface, a technology that allows researchers to simultaneously detect and study thousands of genes (approximately 21000 genes in the human genome) [7]. Currently, two microarray technologies are used, namely complementary DNA (cDNA) and oligonucleotides. Both involve hybridisation, which differs in the placement of the DNA sequences as well as the length of the sequences.

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

cDNA techniques have lengths from a few hundred to several thousand samples. Usually, for most gene expression profiling experiments with cDNA, the mRNA from two different sources (such as diseased cells and normal cells) is extracted, purified, and reverse transcribed in the first strand of cDNA sequences. Fig. 1. presents the most critical steps for this kind of procedure.

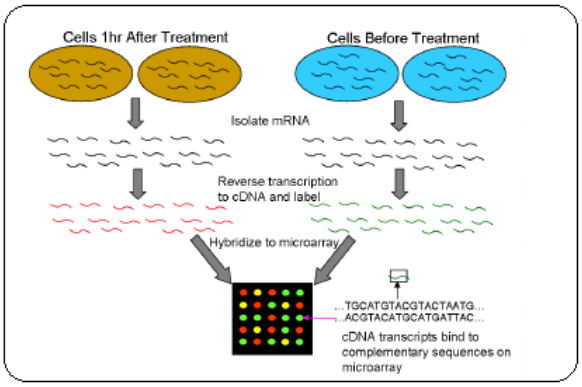

Fig. 1. cDNA technique [7]

The difference between the two channels consists of different hybridisation: mRNA in oligonucleotide microarrays, compared to cDNA in microarrays. Also, the cDNA technique is a two-channel technique, while the RNA technique is a single channel, the types of samples being synthesized directly on the microarray [7].

Several companies were born from the application of microarray analysis technologies in disease prediction, resulting in various gene expression profiling tests [8, 9]. There is a wide range of software tools available for performing biological data analysis and visualization [10]. For example, the statistical programming language R is an essential tool for biological analysis, primarily used for data manipulation, statistical calculation, and graphical display, offering a wide range of statistical techniques (linear, nonlinear modelling, tests, time series analysis, classification, grouping), including a vast collection of data analysis, manipulation and storage tools [11]. It can also be used as a software development tool that can be integrated with Perl and MySQL, facilitating the creation of highly sophisticated bioinformatics software [12]. One of the most popular biological analysis tools is Bioconductor [13], open-source software for analyzing and understanding genomic data implemented in the R program.

With such a large amount of data available to the general public, a bioinformatics researcher needs to possess the specific knowledge and abilities to understand, examine and interpret this data in the most specific way possible. In this paper, we will present a microarray gene analysis approach and extraction of genes of interest.

B. Exploration of gene expression data

Paper [14] proposes an analysis of gene expressions at the genome level in patients with critical diseases, performed with oligonucleotide matrices, in which significant variations of genome expressions are observed between trauma

patients vs. healthy patients. Expectedly, the traumatic lesion induces dramatic changes in gene expression, larger than the analytical noise and interindividual variance. The authors also propose developing a national program for the analysis of gene expression, with increased attention to analytical details.

In paper [15], gene expressions were analyzed using Pearson correlation, entropy, and principal component analysis, resulting in similar groups between T genes, using hierarchical clustering and Ward method in R language. In contrast, in [16], a series of information about databases, statistical models, and graphical programs for the analysis of gene expressions are presented (such as Cytoscape or the analysis of biomolecular networks using Petri models).

The study [17] presents the modelling of gene expression through an autonomous system of delayed differential equations, as well as its software built for gene sequence analysis called ExpertDiscovery, which predicts gene functions using a set of integrated methods for recognizing regulatory and site elements. transcription factor binding.

Article [18] sets out the main steps for building a model to characterize different cancer subtypes, to predict and classify several cancer subtypes based on public genetic expressions. The model is designed to predict gene changes from one subgroup of cancer to another, to identify, test, and subsequently develop new types of treatments for these diseases. The study [19] integrates sets of gene expression profiles from normal and colorectal cancer cells on the same topic. These were analyzed, resulting in differentially expressed genes grouped based on signaling functions and pathways. Of these, 31 were found to be involved in the cell cycle process and identified as candidates that could improve understanding of the cause of colorectal cancer. For complex analysis, the DAVID tool was used, as well as GO (The Gene Ontology Consortium), for the detection of gene ontology categories.

Paper [20] proposes an analysis of primary colon tumour samples, identifying differentially expressed gene networks and grouped into six molecular subtypes associated with distinct phenotype characteristics, molecular alterations, or specific gene signature pathways. The classification was validated using samples explicitly kept for testing, and colon cancer samples from 8 public data sets.

Principal component analysis was applied to reduce the dimensionality of gene expressions data in studies such as [21, 22]. Authors in [22] applied the least-squares method for size reduction and variables selection, while the authors in [23] applied PCA and neural networks to classify cancers using gene expression profiling. This type of analysis was applied to classify tumor samples vs. non-neoplastic tissue.

In [24], scientists use partial least squares analysis as a classification procedure to identify specific genes of interest from many genes, selecting only those highly correlated with the phenotype.

3. Breast cancer microarray analysis approach

The process of knowledge management in the field of gene expression includes several steps, including data preprocessing (data cleaning), data integration (if data from multiple sources are used), transformation, extraction data, associations and correlations, classifications, predictions, groupings, analysis of outliers, the discovery of models, evaluation of models and presentation of knowledge.

In this paper, we analyze gene data from oligonucleotide matrices, called Affymetrix GeneChip, in the R programming language. Our approach is divided into six main steps, the final one bringing a much smaller number of differential expressed genes related to the chosen process.

The main step, to begin with, is to download relevant data from files related to the phenotype of interest (in our case, breast cancer). The biological data GSE48391 was downloaded from the GEO database, meaning files of gene expressions selected from Affymetrix microarrays of breast cancer, i.e. DNA sample files that contain all the data deposited in the gene ontology database, from a series of CEL files in the form of PM and MM intensities. The CEL file stores the results of the intensity calculations on the pixel values in the DAT file. The dimensions of the CEL file are X x Y, (where X are the rows, representing the number of genes on the chip and Y the columns, representing the number of samples of GSM files), the matrix containing different values for the expression level (absolute or relative) of a specific gene in a particular sample or condition. Each column vector contains the results obtained in a certain sample and is called the profile of that condition, and each row vector is a model of the expression of a particular gene in all conditions existing in the GSE file [25].

The second step is removing non-biological elements from the sample set and transforming the intensity values into expression values. There are several methods for data normalization, all techniques making background correction, scaling, and aggregation. In this paper, we propose to use the Robust Multiarray Averaging (RMA) algorithm from the affy package (Fig. 2).

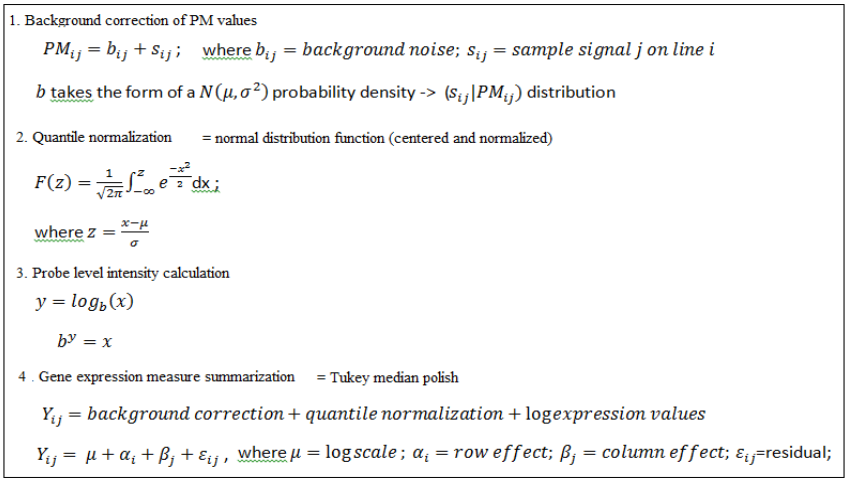

Fig. 2. Main steps for Robust Multiarray Average algorithm

Consisting of 3 steps, background correction of PM values on each matrix separately, normalization, and summary of the measure of gene expression, RMA analysis uses only the information from PM samples to estimate the distribution parameters and returns the estimated signal [26].

One of the significant challenges of microarray data analysis is the much larger number of genes far exceeding the number of existing samples, so an important step that must be done after normalization is the one of dimensionality reduction. To solve the problem of dimensionality, many methods can be used to reduce dimensionality [27, 28, 29], but in our analysis, we present a different approach that begins by selecting the genes expressed in at least 5% of the samples, with a significantly different variance from the median variance of all sets of samples. Thus, we suggest selecting that the 5th decile of each gene expression value be more significant than a specific chosen value and be expressed in at least 5% of the total number of existing samples in the matrix. Then with the help of a chi-square test, using (1) and using a chosen threshold, lower than a particular value, apply a qchisq test [30].

$$\gamma^2 = \frac{(n-1)s^2}{\sigma^2}; \tag{1}$$

where n = total sample size;
s^2 = sample variance (square of standard deviation);
σ^2 = hypothetical population variance;

The deliverables will be a much smaller number of genes, genes that have the chance to characterize better a phenotype that differs from the rest of the population.

A powerful analytical tool to benefit from large sample sizes is grouping. Clustering, the next step to be fulfilled, is an unsupervised method of grouping sets of similar objects based on a particular criterion, usually a series of features whose similarity is defined by a particular distance function [24, 25]. Cluster analysis is strong because it is not based on a class label (such as disease status), allowing the discovery of new relationships between variables. At this stage, based on the hierarchical grouping, the differentially expressed genes will be identified and selected, with a statistically significant p-value, to reject the test hypothesis. We propose that genes differentially expressed between a given cluster and all others be identified using a Welch-type t-test [30]. Based on this criterion (Fig. 3), the list of genes expressed significantly differentially will be established using a threshold value smaller than 0.05.

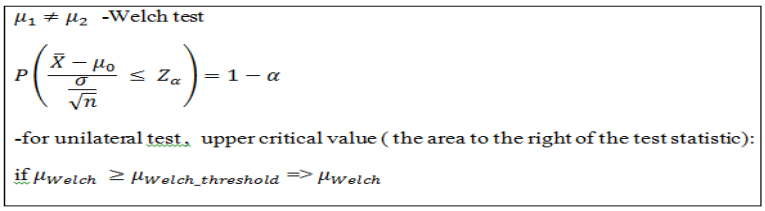

Fig. 3. Welch t-test

Next, based on our reduced set of genes, the Weighted correlation network analysis (WGCNA) package from the R language [31, 32] can be applied to form groups of genes correlated with each other, resulting in relational modules in which the "leading" genes are identified and selected.

The first step for this gene correlation analysis is to identify co-expression networks, defined as weighted and unoriented gene networks. The nodes of a co-expression network correspond to gene expression profiles, and the margins between genes are determined by pairs of correlations between gene expressions. The construction of the weighted correlation network starts from the correlation matrix, by choosing a threshold power, according to the selection criterion, and then the "neighbourhood" TOM matrix, which takes into account the topological similarity and similarities between genes reflected at the level of network topology (2).

$$TOM_{ij} = \frac{\sum_{k=1}^{n} a_{ik} a_{kj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \quad (2)$$

where $a_{ij} = \left|cor(x_i, x_j)\right|^{\beta}$ is the adjacency matrix weighted with x_i şi x_j vectors of expression values for the gene i and j, $k_i = \sum_{k=1}^{n} a_{ik}$ and β threshold parameter that ensures a scale-free network (characterized by a small number of strongly connected hub nodes and a large number of poorly connected nodes).

The second step in WGCNA analysis is to identify those genes in the modules most connected within the module, meaning those genes strongly

correlated with a clinical or phenotypic feature of interest (in our case breast cancer) [32]. To find the hub genes in each module with the highest intramodular connectivity, maximums on the columns have to be calculated for each module separately. Each resulting gene from each module can be annotated using "hgu133plus2.db", “AnnotationDbi.db", and "annotated" packages.

Our approach advances with the igraph package proposal in the R programming language [11], in which each adjacency matrices for each top gene in each module has to be read and edge list files to be created, to create graphs or genetic co-expression networks for each of the correlated genes in each module.

At this point, each one of these genes can be analyzed and validated within a gene database, such as UniProt, Catalogue of Somatic Mutations In Cancer, National Center for Biotechnology Information [2, 4, 33], as genes of interest in multiple cancers. For example, our analysis brought us up a substantial number of genes seen as prognostic markers in breast cancer, including SERPINA1, TFPI2, CPT1A, NKAP, or RAC2, with important roles in the development and regulation of cell growth in the human body.

4. Discussion

The R code contains six parts. The first part is the reading and normalization one, using "BiocManager", "affy" and "GEOquery" packages, while part 2 acts for reducing the size of the genes, using two tests: qchisq test, and Welch t-test. Part 3 provides the gene annotation, while part 4 represents the WGCNA correlations using the "dynamicTreeCut", and "WGCNA" packages. Part 5 means identifying hub genes, while part 6 is building the gene networks of the most correlated genes according to TOM matrix, using the "igraph" package.

There are diverse papers that treat gene expression analysis based on WGCNA [34 - 39], what makes our approach to be considered is taking all steps mentioned into account to reach significant differential expressed genes related to the structure of interest.

Most of the "classic" approaches mentioned above exports the relationships between genes in a visualization platform; our paper comes with a different structure and proposes a visualization of gene graphs using the R package, igraph, based on the correlation values (weight) between genes.

The DAVID tool is often used for evaluation, annotation, and analysis of differentially expressed genes. In our case, gene annotation was done using multiple annotation packages within the R programming language, identifying both gene symbol and Entrez ID, a unique identifier used for integrating different types of information from several sources into a single one.

Fig. 4 presents a comparison of seven scientific papers, including ours, between all the identified hub genes, annotated and analyzed, with different databases.

Paper	Nb of modules	Breast cancer gene hub	Annotation	Visualisation	Analysis databases
34	18	5 (CCNB2, FBXO5, KIF4A, MCM10 şi TPX2)	DAVID	x	TCGA
35	8	12 (AURKA, BUB1B, CCNB2, CDK1, CDT1, HJURP, KIF20A, KIF2C, KIF4A, MELK, TPX2, UBE2C)	DAVID	x	TCGA, STRING
36	17	6 (EPCAM, MELK, KRT8, KRT19, KPNA2 şi ECT2)	DAVID	Cytoscape	GO, KEGG, Oncomine
37	11	12 (APC, ATRX, CHD1, CHD9, COL4A3BP, DCP2, DMXL1, KIAA1033, RAPGEF6, TRIM23, TTC37 ZFYVE16)	GO	Cytoscape	TCGA, KEGG
38	4	8 (TXN, ANXA2, TPM4, LOXL2, TPRN, ADCY6, TUBA1C, CMIP)	FunRich	R tidygraph	TCGA, KEGG, GSEA
39	5	5 (FABP7, CXCL3, LOC284578, CAPN6, NRG2)	Metascape	x	TGGA, GEPIA, HPA
our paper	13	10 (SERPINA1, TFPI2, RAC2, ARID5A, NKAP, TBL1X, CPT1A, EPRS1, FLT3, CSN1S1)	R packages	R igraph	COSMIC, UniProt, canSAR, Protein Atlas

Fig. 4. Papers comparison on breast cancer gene hubes

Our analysis approach started from 54675 genes and 81 samples, after dimensionality reduction step (Chi-square test) reached 19847 genes, after grouping step (Welch test) reached 7189 genes, and after the Weighted Correlation Network Analysis, the genes were divided into 13 modules. Each one of the modules can be further analysed in terms of relevance to our study of interest. For example, from the 10 identified hub genes, five were identified in the gray module, two in the purple module, and one in the greenyellow, turquoise, and yellow module, in other words, our analysis found genes of interest in breast cancer, validated within gene databases.

In plus, two GSE files validate our analysis approach presented in the paper by identifying essential genes, which play a significant role in cancer. Following our proposal for gene expression analysis, we downloaded GSE36295 [40], which contains RNA isolated from surgically excised breast cancer tissues, purified, labelled, and hybridised with Affymetrix Human Gene 1.0 ST Array. After finishing all filtering parts, 13 genes were selected from the module with the most correlated genes: SKAP2, ITGB5, MYCT1, PODNL1, TEK, SDR42E1, TFPI, JPH1, PAPSS2, MANEAL, PPIC, COBLL1, EMCN. After they had been analyzed and validated within the Catalogue Of Somatic Mutations In Cancer database, we learned that they either regulates the immune system, either cancer-related genes or prognostic markers in breast cancer.

Another data file, GSE102907 containing messenger RNA extracted from the primary tumor of breast cancer patients, hybridised and scanned with the Affymetrix Human Genome GeneChip U133 Plus 2.0 matrix, was downloaded from NCBI [41]. After proceeding with our proposal, different genes were selected from the module containing the most correlated genes: MIPOL1, CDC20B, C7orf57, MPV17L, ZBTB18, CYP4F8, MYBPC1, KITLG, FAM110C, CSTF3, CROT, ARMC3, most of them being cancer disease-related gene.

5. Conclusions

Microarray experiments are designed to achieve one or more goals, such as identification of genes whose expression is correlated with a specific phenotypic trait (response to treatment, causes leading to cancer), identification of genes involved in regulatory and mediating networks for certain biological phenomena or molecular markers that can be used as tools for diagnosing and predicting diseases or even as predictors of clinical outcomes, or discovering possible molecular targets for drug development;

The general objective of any data mining task is to find certain patterns or trends that help to understand the data better. In parallel, the objective of this paper was to present a framework approach to select and identify the main genes that participate in cell cycle control and may undergo mutations in cancer.

As advantages of this solution compared to other "classic" approaches we mention the fact that our entire analysis until the validation of the modules is performed entirely in the R programming language, without other auxiliary platforms (DAVID, Cytoscape, GOrilla), R having the advantage of a high-performance environment analysis that meets the interests of several applications together. Regarding the analysis of genes of interest from the modules, the databases used for the validation of biomarkers are some of the most used in this branch.

The originality of this paper is an integrated R-language analysis [42], different from the usual ones. The analysis begins with double filtering of the gene set, at the end of which it resulted in a small number of differentially expressed genes, divided into modules, using the WGCNA correlation analysis. Based on the correlation link of the top genes in each module, gene co-expression networks were created using the igraph package, from which, using multiple public databases, such as COSMIC, UniProt, canSAR, or Protein Atlas, genes can be further interpreted to see how they relate to the proliferation of altered breast cells in the body.

R E F E R E N C E S

[1]. *A. Pavlopoulou, D.A. Spandidos, I. Michalopoulos*, "Human cancer databases (Review)", Oncol Rep, 33(1): 3–18, doi: 10.3892/or.2014.3579, 2015.
[2]. https://www.ncbi.nlm.nih.gov/, accessed on 2021-11-11.
[3]. https://www.rcsb.org/, accessed on 2021-11-11.
[4]. https://www.uniprot.org/, accessed on 2021-11-11.
[5]. https://academic.oup.com/nar, accessed on 2021-11-11.
[6]. *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.
[7]. *J. Ernst*, "Computational Methods for Analyzing and Modeling Gene Regulation Dynamics", School of Computer Science, Machine Learning Department, CMU-ML-08-110, 2008.

[8]. https://www.breastcancer.org/symptoms/testing/types/oncotype_dx, accessed on 2021-11-11.
[9]. https://agendia.com/mammaprint/, accessed on 2021-11-11.
[10]. *R. Wunschiers,* "Computational Biology. A practical introduction to BioData Processing and Analysis with Linux, MySQL, and R", Second Edition, Springer. DOI 10.1007/978-3-642-34749-8, 2013.
[11]. r-project.org, accessed on 2021-12-11.
[12]. *C. Bessant, D. Oakley, I. Shadforth,* "Building Bioinformatics Solutions woth Perl, R, and SQL", Second Edition, Oxford University Press, 2014.
[13]. https://www.bioconductor.org/, accessed on 2021-12-11.
[14]. *J. Perren Cobb, Michael N. Mindrinos, Carol Miller-Graziano, Steve E. Calvano, Henry V. Baker et all,* "Application of genome-wide expression analysis to human health and disease", PNAS, 102 (13) 4801-4806, 2005.
[15]. *O.Simeoni, V.Piras, M.Tomita, K. Selvarajoo,* "Tracking global gene expression responses in T cell differentiation", Gene 569, 259-266, 2015.
[16]. *C.W. Sensen,* "Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical & Legal Issues", Vol1, Wiley-VCH, ISBN: 3-527-31348-6, 2005.
[17]. *Kolchanov, R. Hofestaedt, L. Milanesi,* "Bioinformatics of Genome regulation and structure II", ISBN-10: 0-387-29450-3, Springer, 2006.
[18]. *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, 2015.
[19]. *Y. Guo, Y. Bao, M. Ma, W. Yang,* "Identification of Key Candidate Genes and Pathways in Colorectal Cancer by Integrated Bioinformatical Analysis", International Journal of Molecular Science, 18(4): 722, doi: 10.3390/ijms18040722, 2017.
[20]. *L. Marisa, A. de Reyniès, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, et al,* "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value", PLoS Med;10(5):e1001453, doi: 10.1371/journal.pmed.1001453, 2013.
[21]. *O'Reilly, C. Gibas, P. Jambeck,* "Developing Bioinformatics Computer Skills. An introduction to Software Tools for Biological Applications", First Edition, ISBN: 1-56592-664-1, 2001.
[22]. *T. Mehmood, H. Martens, S. Saebo, J. Warringer, L. Snipen,* "A Partial Least Squares based algorithm for parsimonious variable selection", Algorithms for Molecular Biology, 6:27, 2011.
[23]. *J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer,* "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Med. 7 (6), 673–679, 2001.
[24]. *T. Brown,* "Introduction in genetics. A molecular approach", Garland Science, ISBN 978-0-8153-6509-9, 2012.
[25]. *P.Baldi, S. Brunak*, "Bioinformatics. The machine learning approach", Second Edition, MIT Press, 2001.
[26]. https://cbdm.uni-mainz.de/files/2016/02/GE_microarrays.pdf, accessed on 2021-09-11.
[27]. *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.
[28]. *P.P. Sinha*, "Bioinformatics with R Cookbook. Quick answers to common problems", Packt Publishing, ISBN 978-1-78328-313-2, 2014.

[29]. *M.L. Raymer, W.F. Punch, E.D. Goodman, LA. Kuhn, A.K. Jain,* "Dimensionality reduction using genetic algorithms", IEEE Transactions on Evolutionary Compututation, 4(2) : 164-171, 2000.
[30]. *L.V.Boiculese, G. Dimitriu, M. Moscalu*, "Elemente de biostatistica. Analiza statistica a datelor biologice", Editura PIM, ISBN: 978-973-716-523-7, 2007.
[31]. *P. Langfelder, S. Horvath,* "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics volume 9, Article number: 559 (2008).
[32]. *S. Horvath, J. Dong,* "Geometric Interpretation of Gene Coexpression Network Analysis", Plos Computational Biology, https://doi.org/10.1371/journal.pcbi.1000117, 2008.
[33]. https://cancer.sanger.ac.uk/cosmic/, accessed on 2021-10-11.
[34]. *J. Tang, D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong and 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.
[35]. *G. Shi, Z. Shen, Y. Liu, W. Yin,* "Identifying Biomarkers to Predict the Progression and Prognosis of Breast Cancer by Weighted Gene Co-expression Network Analysis", Front. Genet., https://doi.org/10.3389/fgene.2020.597888, 2020.
[36]. *J. Qiu, Z. Du, Y. Wang, Y. Zhou, Y. Zhang, Y. Xie, Q. Lv*, "Weighted gene co-expression network analysis reveals modules and hub genes associated with the development of breast cancer", Medicine (Baltimore);98(6):e14345. doi: 10.1097/MD.0000000000014345, 2019.
[37]. *L. Lan, B. Xu, Q. Chen, J. Jiang, Y. Shen,* "Weighted correlation network analysis of triple negative breast cancer progression: Identifying specific modules and hub genes based on the GEO and TCGA database", Oncology letters, Pages: 1207-1217, https://doi.org/10.3892/ol.2019.10407, 2019.
[38]. *C. C.N. Wang, C.Y. Li, J.-H. Cai, P.C.-Y. Sheu, J.J.P. Tsai et al.,* "Identification of Prognostic Candidate Genes in Breast Cancer by Integrated Bioinformatic Analysis", J. Clin. Med, 8(8), 1160; https://doi.org/10.3390/jcm8081160, 2019.
[39]. *J. Wu, X.-J. Liu, J.-N. Hu, X.-H. Liao, F.F. Lin*, "Transcriptomics and Prognosis Analysis to Identify Critical Biomarkers in Invasive Breast Carcinoma", Technology in Cancer Research & Treatment, https://doi.org/10.1177/1533033820957011, 2020.
[40]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36295, accessed on 2021-12-11.
[41]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102907, accessed on 2021-12-11.
[42]. https://github.com/irishptr/gene_expression_analysis.