MODELLING BREAST CANCER GENE EXPRESSION USING BAYESIAN NETWORKS

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

This paper addresses microarray analysis, through an integrated technique of interpreting gene expressions from modified breast tissue in order 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 to the appearance of this neoplasm. The approach described in this paper is helpful in discovering and understanding as fully as possible the causal relationships between genes identified as being differentially expressed from a biological data set.

Keywords: microarray analysis, gene, modelling, Bayesian networks

1. Introduction

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 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 main objective of this article was to identify genes whose expression is correlated with a specific 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 helpful in discovering and understanding as fully as possible the causal relationships between genes identified as being differentially expressed from a biological data set.

[1] PhD Student, 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

2. Gene expression profiling techniques

Microarray-based gene expression has propelled our knowledge of molecular biology [1]. First of all, they have become a widely used technique to study the dynamics of biological processes, being miniaturized laboratories for studying gene expression [2]. 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 [3]. We can obtain responses to biological components that interact with each other, microarray analysis having various applications in the medical field, starting from the characterization of tumours to the evolution or changes of diseases or symptoms over time or response to drugs and identification of new treatments [4]. 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, such as comparing a young transcriptome with an old one [5], revealing changes in molecular pathways. Even a partial understanding of the information available can provide value and clues. For example, co-expression of new genes may provide functions for many genes for which information is unavailable.

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 many genes (approximately 21,000 genes in the human genome) [6]. Microarray experiments and information sequence analysis processes are designed to achieve one or more goals, such as:

- ➔ identification of genes whose expression is correlated with a particular phenotypic trait;
- ➔ identification of genes involved in regulatory and mediating networks for certain biological phenomena;
- ➔ identification of molecular markers that can be used as tools for diagnosing and predicting diseases or as predictors of clinical outcomes;
- ➔ discovering possible molecular targets for drug development;

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 makes data processing a cumbersome task. This task can be further complicated by the inevitable batch effects generated when combining different data sets or the noise present in all-time series expression experiments. 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.

3. Approach

A. Microarray analysis in R software

Data from gene expressions can lead to complex applications such as discovering 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, a bioinformatics analyst needs 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 and model gene data from oligonucleotide matrices from faces called Affymetrix GeneChip [7] in the R programming language [8]. The data are implemented in Affy matrices, the expression of each gene being measured by comparing the hybridization of the molecular RNA of the sample with a set of PM and MM probes. The first sample type in each pair is called the perfect match (PM) and is taken from the gene sequence. The second type of sample is called mismatch (MM), which measures background noise and is created by changing the 13th gene in PM.

Initially, the biological data was downloaded from the Omnibus Gene Expression (GEO) database, hosted by the National Center for Biotechnology Information (NCBI) [9]. From this public repository of biological files, the GSE48391 file of gene expressions on the Affymetrix microarray faces of breast cancer was selected.

The first steps used in our approach were fully explained in [10]. After microarray data normalization, solving the issue of high dimensionality and hierarchical grouping, we used the R Weighted Correlation Network Analysis (WGCNA) package [11], which forms groups of genes correlated with each other resulting in relational modules (Figure 1).

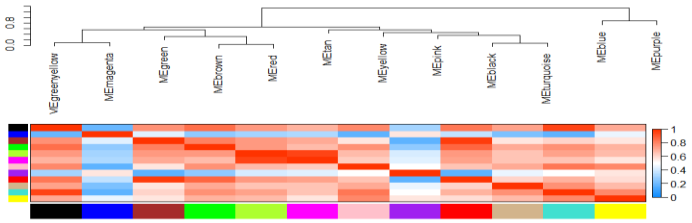

Fig. 1. Gene co-expression modules grouped according to Euclidean distance

From 13 modules, and according to the theory of the hclust function [12], the group with the most correlated gene (in our case the greenyellow module) was chosen for further analysis.

Using the igraph package from the R programming language, the adjacency matrice was used to generate correlation scores between nodes (hub

genes). By sorting the genes according to the weight value, we selected the first 30 correlations in the order of the decreasing weight values, resulting in a particular graph for the chosen module.

Using the gene co-expression network (GCN) of the module with the most correlated genes, an undirected graph with a single directional path was created to exemplify the possible alteration pathways leading to the formation of cancer cells (Figure 2).

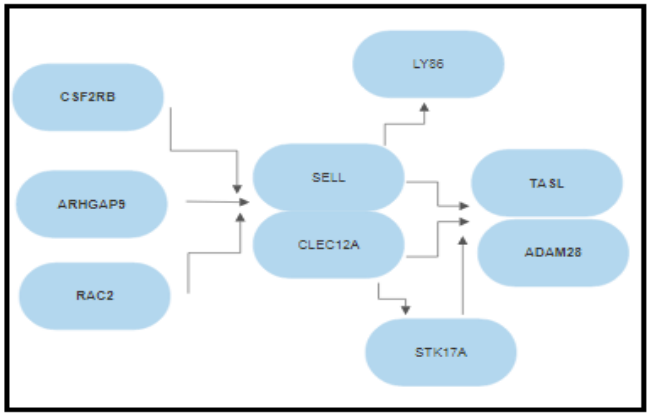

Fig. 2. Schematic reproduction of the gene system within the module chosen to be modeled

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

Next, with the help of Bayesian networks and inference, we analyzed the effects that possible mutations or polymorphisms in each gene alter or not the role of the ADAM28 gene in the module chosen as representative of the tumour progression of breast cancer.

B. Analysis of gene expressions based on Bayesian networks

Bayesian networks can be an approach to the analysis of gene expressions and their patterns by statistical examination of dependencies and conditional (in) dependencies of data, becoming an essential part in the genetic analysis of data, being used to deduce causal relationships [13, 14]. These are a class of probabilistic models used to model reasoning under uncertainty. Each Bayesian network has two components:

- ➔ Qualitative component, which uses the language of graphs, suggests a set of (in) dependency relationships between domain variables (parent-child dependencies);
- ➔ Quantitative component, corresponding to the probabilistic modelling of uncertainties, using probability theory (we attribute to dependency relations - marginal probabilities of all nodes without parents and conditional probability distributions to the other nodes giving the parent nodes);

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 [15]. Compared to the diagram in Figure 3, the SELL gene with the related offspring was eliminated. According to research and studies conducted so far, the rest of the genes were actively involved (together or separately) in breast cancer neoplasms. Thus, the process in Figure 3. begins with three genes, of which RAC2 is part of the Ras family, being a proto-oncogene with a role in regulating cellular responses, such as apoptotic and epithelial cell processes, CSF2RB - receptor for a growth factor that induces differentiation and proliferation in the spinal cord bone, and ARHGAP9 a suppressor gene, which if mutated, leads to deletion of the p53 gene, thus promoting the invasion of fibroblast cells into cells and tissues [16]. CLEC12A encodes essential proteins involved in cell signaling and immune response (some of which are part of the 'killer genes' region), STK17A is a suppressor gene, a member of cell apoptosis and a target of the p53 gene in case of mutations [17], and ADAM28 an oncogene with immune cell binding, overregulated in specific cancer cells [18], mutations in gene expression being linked to metastatic dissemination [19].

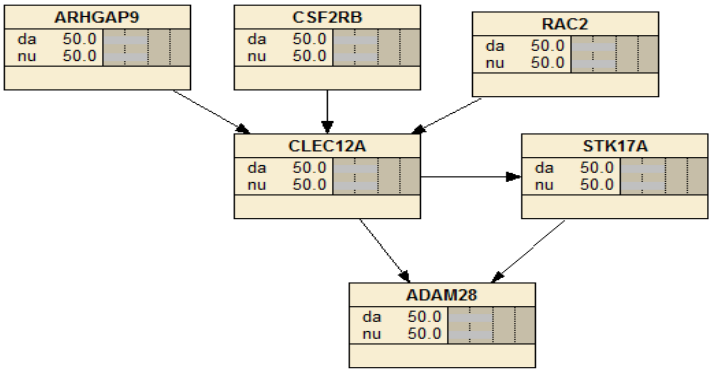

Fig. 3. Targeted acyclic graph with genes susceptible to breast cancer

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, one needs the quantitative component, i.e. both the marginal probabilities of the variables attached to the nodes and the conditional probability tables associated with the nodes. In the first phase, we assumed that each gene had equal marginal probabilities, after which we modelled the network for different probabilities to analyze which are the primary mutations in genes that can lead to changes in the structure of the ADAM28 gene. ADAM28, referring to all the other genes in the process.

Starting from the compound probability formula [14], the Markov causality condition states that a Bayesian network uniquely defines a factorization of the compound probability distribution (where "children" depend only on "parents"). In our case, the formula is:

$$P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28) = P(ARHGAP9)P(CSF2RB)P(RAC2)P(CLEC12A \mid ARHGAP9, CSF2RB, RAC2) P(STK17A \mid CLEC12A)P(ADAM28 \mid CLEC12A, STK17A) \quad (1)$$

The probability of mutations in the ADAM28 gene, given the other genes, can be written mathematically using the Markov causality condition:

$$P(ADAM28 \mid ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A) = P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)/ (\sum_{ADAM28} P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)) = P(ADAM28 \mid CLEC12A, STK17A) \quad (2)$$

Thus, the last part of the gene analysis assumed the modelling of the genes in the chosen module to have the most significant relevance to the established condition, with the help of a Bayesian network that uniquely defines a composite probability distribution over all variables in the network. Knowing the composite probability of the network, it is easy to calculate the marginal or composite probability of any variables in the network. This has also been done in this case with the help of the Bayesian inference, which is very flexible, allowing the introduction of records on any node and updating the trusts for any of the other nodes.

Depending on the existence of a particular type of evidence or information (diagnostic tests, ultrasounds, blood sampling) on the genes, we can provide additional information about their changes in the chosen biological process. Thus,

if there is no reliable information from the physician or expert, Bayesian network modelling 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 specific information is known, the a posteriori probabilities of the nodes are calculated.

In the analysis of the influence of specific information on each node in the system, some 7 cases of probabilistic inference of causal or predictive reasoning resulted:

a. If the information about 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 "parent" RAC2, CSF2RB, ARHGAP9 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 accurate, 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 brought about the “parent” genes RAC2, CSF2RB, ARHGAP9 does not change the trust 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 brought about 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 provided on the "parent" genes RAC2, CSF2RB, ARHGAP9 does not change the confidence in the CLEC12A and STK17A genes;
- ➔ any (certain) information provided on the "parent" genes RAC2, CSF2RB, ARHGAP9 does not change the confidence in the ADAM28 gene, so we can state that the initial genes and the final gene are conditionally independent.

4. Verification elements and limitations. Discussions

The main object of the paper was to identify genes whose expression is correlated with a specific phenotypic trait, in our case, breast cancer and which can be used as tools for possible pathways through which mutations of differentially expressed genes lead to the appearance of this neoplasm. Thus, Bayesian networks were chosen for modelling the gene co-expression network, which was helpful in modelling reasoning under uncertainty. In addition, they have an intuitive and flexible language for representing the dependencies and independence 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 to analyze (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. The reasoning analyzed in this paper is causal or predictive, in which predictive inference takes place in a causal sense, namely, we were able to answer questions such as: giving the cause (s), what is the chance of producing the effect? In our case, what is the probability that a patient will have mutations in the ADAM28 gene knowing that the level of mutations in the upper genes is high?

For our analysis, clinical information on breast cancer from the Catalog of Somatic Mutations in Cancer (COSMIC) [20] was used to validate genes we identified and analyzed as related to the phenotype of interest in the module chosen to be modelled.

One independent data set was used to validate the primary gene in our analysis. To do this, we downloaded the GSE102907 file [21], a file containing messenger RNA extracted from the primary tumour of breast cancer patients, hybridized and scanned with the Affymetrix Human Genome matrix GeneChip U133 Plus 2.0, the ADAM28 gene being found as a differentially expressed gene.

Using the Kaplan Meier plotter [22], all genes analyzed in the chosen gene network were identified in this application as biological biomarkers (which can provide information about a person's health such as the presence or stage of a disease, a physiological change in the body, a reaction to a treatment, a psychological state).

Many papers state that the ADAM28 gene is overexpressed in several cancers, including breast cancer [23], but none of the microarrays analysed in publicly available works 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 (whose presence and change in concentration is correlated with the development of tumours), 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.

As a limitation of the analysis and identification of possible pathways of mutations in genes susceptible to control and regulate cell cycle progression and apoptosis, we can specify that the probabilities of nodes were chosen randomly, real data about these genes has a much more substantial 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 realistic analysis. Moreover, the availability of papers attesting to the links between the genes analyzed in this paper 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 help a better understanding of the system.

5. Conclusions

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 double filtering of the gene set with the help of two statistical tests: Chi-square and Welch t-test, the end of which resulted in a small number of differentially expressed genes of statistical significance. Based on these genes, using the WGCNA correlation analysis, thirteen gene modules were identified. Based on the correlation link of the top 30 genes in each module, gene co-expression networks were created using the igraph package. The most biologically relevant gene co-expression module (of all the ones we identified) in breast cancer was selected from all the resulting networks. 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. Developed by Norsys [24], the environment has the advantage of an intuitive user interface, offering flexibility in defining the relationships between variables and displaying the results of inference. With the help of Bayesian inference, the effects that any mutations in one or more genes in a module chosen as having biological relevance in breast tissue neoplasia may have on the gene chosen as representative of breast cancer progression have been analyzed, namely ADAM28, the overexpressed gene in human breast carcinomas, its expression being linked to tumor progression and metastatic dissemination.

As future prospects, the first thing we set out to test is the possibility of intervening in the causal process. 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 other nodes.

This methodology can be applied to several types of cancer as a general framework, so another idea would be the comparison of genes specific to various types of cancer, to see what are the similarities and differences between them.

Moreover, another future perspective would be the identification of differentially expressed genes from an altered material compared to a genetically healthy one and their comparison to identify the differences between the two genetic materials.

Of great interest would also be the exact finding of amino acids in the differentially expressed genes that change when mutations occur. In which areas they occur, what types of areas are (critical or not), if those mutations are expressed or remain silent, or what is the rate of their mutation, all these are other key questions whose answers would help a better understanding of how genes communicate in the human body.

R E F E R E N C E S

[1]. *J.M. Keith*, "Bioinformatics Vol I. Data, Sequence Analysis and Evolution", Methods in Molecular Biology, Humana Press, ISBN: 978-1-58829-707-5, 2008.
[2]. *J. Ernst*, "Computational Methods for Analyzing and Modeling Gene Regulation Dynamics", School of Computer Science, Machine Learning Department, CMU-ML-08-110, 2008.
[3]. *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.
[4]. *P. Baldi, S. Brunak*, "Bioinformatics. The machine learning approach", Second Edition, MIT Press, 2001.
[5]. *M.R. Speicher, S.E. Antonarakis, A.G. Motulsky*, "Vogel and Motulsky's Human Genetics. Problems and Approaches 4th Edition", Springer, 2010.
[6]. *T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi*, "Hidden Markov models for cancer classification using gene expression profiles", Information Sciences 316, 293-307, 2015.
[7]. *G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger*, "The analysis of Gene Expression Data. Methods and Software", Springer, ISBN 0-387-95577-1, 2003.
[8]. The R Project for Statistical Computing. https://www.r-project.org/, accessed on 2022-03-11.
[9]. Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/, accessed on 2022-01-19.
[10]. *I.-O. Lixandru-Petre, C. Buiu*, "An integrated breast cancer microarray analysis approach", U.P.B. Sci. Bull., Series C, Vol. 84, Iss. 2, ISSN 2286-3540, 2022.
[11]. *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.
[12]. Hierarchical Clustering function. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust, accessed on 2022-02-21.
[13]. *I. Mihu*, "Tehnici de decizie şi diagnoză", Editura Universitară, 2008.
[14]. *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.
[15]. *A. Pavlopoulou, D.A. Spandidos, I. Michalopoulos*, "Human cancer databases (Review)", Oncol Rep, 33(1): 3–18, doi: 10.3892/or.2014.3579, 2015.
[16]. *F. Guo, Y. Zheng*, "Rho family GTPases cooperate with p53 deletion to promote primary mouse embryonic fibroblast cell invasion", Oncogene volume 23, pages 5577–5585, 2004.
[17]. *W. Wei, C. Liu*, "Prognostic and predictive roles of microRNA-411 and its target STK17A in evaluating radiotherapy efficacy and their effects on cell migration and invasion via the p53

signaling pathway in cervical cancer", Mol Med Rep, 21(1): 267–281, doi: 10.3892/mmr.2019.10826, 2020.
[18]. *Y. Mitsui, S. Mochizuki, T. Kodama, M. Shimoda, T. Ohtsuka, T. Shiomi, M. Chijiiwa, T. Ikeda, M. Kitajima and 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): 9913-9920, 2006.
[19]. *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.
[20]. Catalogue of Somatic Mutations in Cancer. Expression, https://cancer.sanger.ac.uk/cosmic/help/expression, accessed on 2022-03-21.
[21]. Expression profiling by array. GSE102907. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102907, accessed on 2022-01-11.
[22]. https://kmplot.com/analysis/, accessed on 2022-02-11.
[23]. *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.
[24]. Netica software. https://www.norsys.com/, accessed on 2022-01-13.