Current State of the Problem of Gene Expression Data Processing and Extraction to Solve the Reverse Engineering Tasks in the Field of Bioinformatics

Sergii Babichev [a,b], Mykhailo Yasinskyi[c], Lyudmyla Yasinska-Damri[c], Yurii Ratushniak [c] and Volodymyr Lytvynenko[d]

[a] *Jan Evangelista Purkyne University in Usti nad Labem, Ceske mladeze, 8, Usti nad Labem, 40096, Czech Republic*
[b] *Kherson State University, University str., 27, Kherson, 73003, Ukraine*
[c] *Ukrainian Academy of Printing, Pid Goloskom str., 19, Lviv, 79020, Ukraine*
[d] *Kherson National Technical University, Berislavske shose, 24, Kherson, 73008, Ukraine*

Abstract

The review presents the current state of the problem of gene expression data processing and extraction for purpose of both the following gene regulatory network reconstruction or the diagnostic systems of complex disease creation. We have described the stepwise procedure of gene expression data processing from initial gene expression values matrix formation to extraction of the most informative gene expression profiles in terms of their separate ability to identify the investigated object. The experimental foundations for our review are arrays of gene expression obtained as a result of both DNA microarray experiments or RNA molecules sequencing methods. The presented analysis considers not only the current state of research in this subject area but and the authors' experience in this direction with the allocation of unsolved parts of the general problem.

Keywords

RNA molecules sequencing, gene expression data, DNA microchip experiment, gene expression profiling processing and extraction, exploratory analysis, clustering, classification, diagnostics, bioinformatics.

1. Introduction

This research is focused on gene expression profiling processing and extraction of genes which can allow distinguishing the investigated objects (healthy or ill, type of disease, etc.) with the higher level of resolvability considering the type of decease for the creation or improve the techniques of both the reverse engineering task solving and effective patients' state diagnostic systems create. This problem is nowadays one of the current areas of bioinformatics [1-4].

The experimental foundation of the research are arrays of gene expressions obtained as a result of both DNA microarray experiments [5,6] or RNA molecules sequencing methods [7,8]. Expression of gene values, in this case, indicates a degree of gene activity. This value is rateable to the gene quantity that satisfies the corresponding type of protein in the biological object.

Gene expression profiling is a vector of expression values certain for unlike biological probes or under the various conditions of the test carried out. Reverse engineering task and further modelling

IntelITSIS’2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine

EMAIL: sergii.babichev@ujep.cz (S. Babichev); yasinskyimf@ukr.net (M. Yasinskyi); Lm.yasinska@gmail.com (L. Yasinska-Damri) ; yurii.ratushniak@gmail.com (Y. Ratushniak); immun56@gmail.com (V. Lytvynenko)
ORCID: 0000-0001-6797-1467 (S. Babichev); 0000-0003-2893-0464 (M. Yasinskyi); 0000-0002-8629-8658 (L. Yasinska-Damri) ; 0000-0003-2995-8610 (Y. Ratushniak); 0000-0002-1536-5542 (V. Lytvynenko)

the reconstructed of the qualitative gene networks form the foundations for exploration and survey of both the nature of network's elements interconnections and their influences on the dynamical possibilities of biology objects.

The complicacy of reverse engineering problem solving is defined by the following: the applied experimental data does not allow defining exactly the network topology on the one hand, and the design of network nodes interconnection on the other one. Moreover, a huge amount of nodes (genes, metabolites, etc) intricate the explanation of the network nodes interconnections. In this case, we need to research: experimental data processing to establish the best ways of gene expression array forming; profiles of gene expression values extraction to allocate the most informative genes considering the level of their separate ability using quantitative criteria.

A qualitatively reconstructed gene regulatory network (GRN) allows exploring the model of the biological system functioning at the genetic level. These facts form the provisos for both making new active medicines and the grown of the techniques of both early diagnostics and effective treatment of intricate diseases. Presented hereinbefore indicate the urgency of the study in this direction.

2. Formal problem statement

Fig. 1 shows a stepwise process of gene expression profiling handling for solving the reverse gene engineering task.

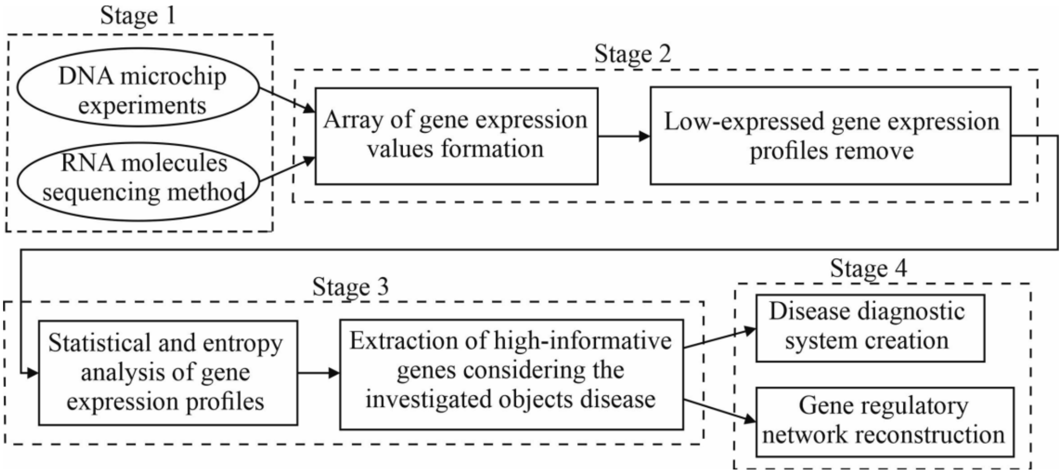
Stage 1
DNA microchip experiments
RNA molecules sequencing method
Stage 2
Array of gene expression values formation
Low-expressed gene expression profiles remove
Stage 3
Statistical and entropy analysis of gene expression profiles
Extraction of high-informative genes considering the investigated objects disease
Stage 4
Disease diagnostic system creation
Gene regulatory network reconstruction

Figure 1: A stepwise process of gene expression profiling handling for solving the reverse gene engineering task

As we can see from Fig. 1, the execution of this manipulation guesses four stages. At the first stage, it is necessary to select experimental data considering the type of experiment performing. At the first stage, it is necessary to select experimental data considering the type of experiment performing. Two techniques are used mainly nowadays to generate the gene expression profiling data: DNA microarray tests and RNA-molecules sequencing process. In the first type of experiments, the data are introduced as the CEL files. Each of them contains the result of the experimental research for one of the executed probes. Thus, the amount of files is equal to the amount of executed tests. Four steps are performed to transform the received test data into a matrix of gene expression profiling: background correction, normalize, PM correction, and summarize. Each of these steps can be executed using diverse tools that are available in the Bioconductor package. As a result, we receive the matrix of gene expression profiling where rows and columns are genes and examined objects respectively.

In the case of another type of experiment apply, we have as a result the list of arrays where each of them contains appropriate data from data annotation to counts of appropriate genes for examined objects. Thus, we can allocate the matrix of genes count for executed objects and further, transform them into expression values directly. We would like to remark that the RNA molecules sequencing test is much more accurate than the DNA microarray technique. However, it is more expensive therefore nowadays two techniques are applied to create the experimental data.

The second stage involves forming the matrix of gene expression profiling and removing zero-expressed and lowly-expressed genes. Initially, the dataset contains about 50000 genes, half of them are zero-expressed for all objects, and for this reason, these objects can be deleted without lost useful information. Then, we should extract lowly-expressed genes that do not allow us to differentiate the objects with a dissimilar state. The amount of genes is decreased at this step to approximately 10000 ones.

The next stage assumes using both the data mining and machine learning techniques for high informative genes extraction considering both the type and health state of the patients' disease. These genes are used at the fourth stage to create the diagnostic system or to reconstruct the gene regulatory network.

Further, we will describe the techniques of each of the stages performing with the allocation of the existed difficulties and unsolved parts of the general problem.

3. Particularities of the experimental data formation

From the presented hereinbefore we can conclude that two techniques are used in most cases to generate the gene expression profiling data nowadays: DNA microarray tests and RNA molecules sequencing experiments [9-17]. The obtained experimental data are freely available at various web resources such as Array Express [18] etc.

Figure 2 shows the step-by-step procedure of the DNA microchip experiment performing in the case of cancer cell use [19]. As it can be seen, two types of cells are used for microchip creation: healthy cells in green color and cancer cells in red color. After the hybridization procedure, the surface of different color spots has been obtained, where the appropriate color existence and light intensity determine the existence and quantity of the corresponding gene. Scanning the microchip allows us to convert the color matrix into a numeric matrix where each of the values is proportional to the amount of the corresponding type of gene.

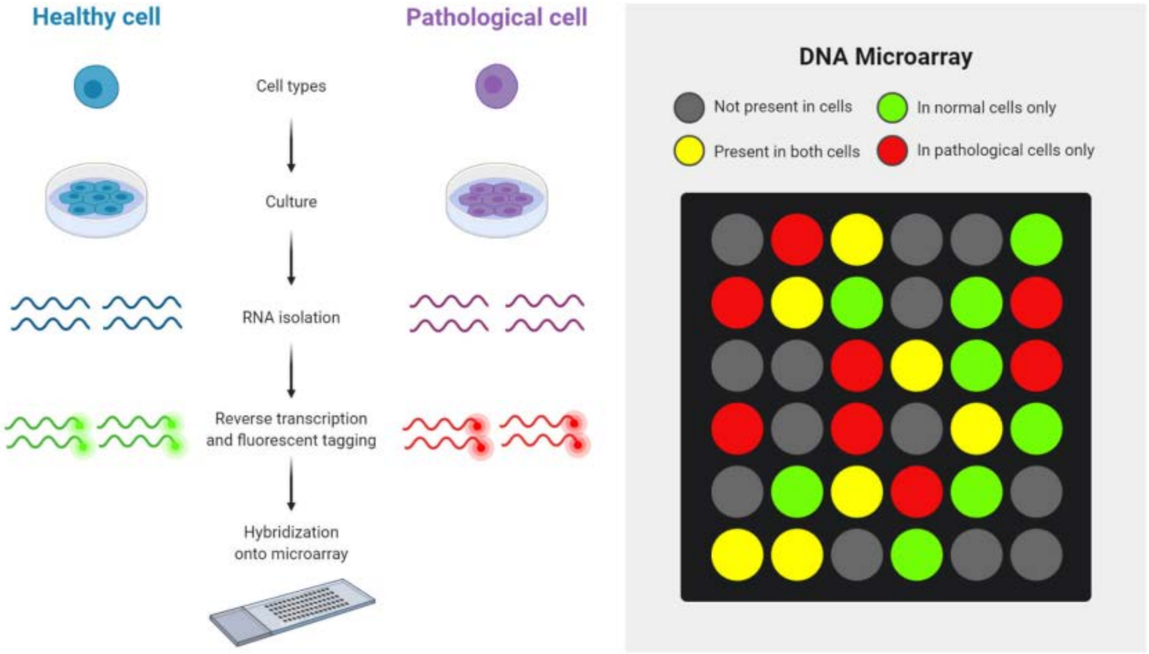

Figure 2: A stepwise procedure of DNA microchip experiment performing [19]

In [9-11] the authors have shown that converting the obtained numeric matrix into a gene expression array involves four steps: background correction, normalization, PM correction, and summarization. Each of the stages assumes using various methods. Selection of the optimal combination of these methods using quantitative quality criteria is nowadays one of the unsolved tasks. The use of various combinations of methods leads to significantly different results. This fact certainly has an impact on the experimental error.

In [20], the authors have considered the way DNA microarray data handling using the Shannon entropy measure assessment based on the James-Stein shrinkage estimator [21].

Figure 3 shows the stepwise procedure of RNA molecules sequencing method implementation [22].

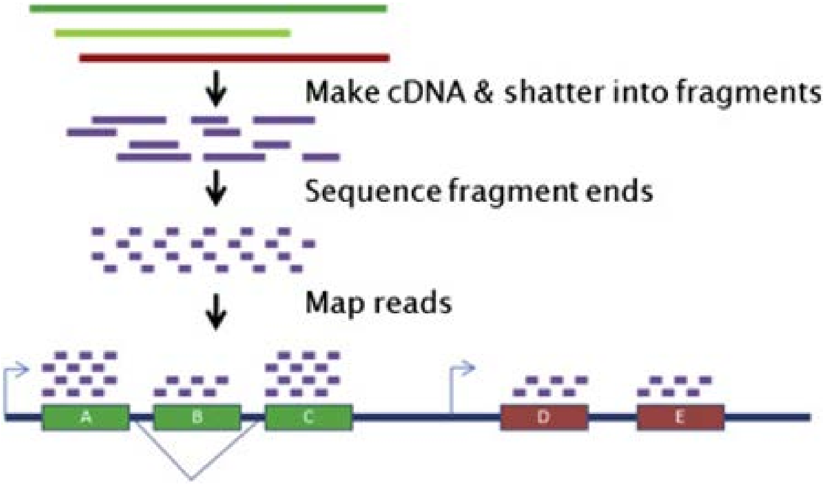

Figure 3: A stepwise procedure of RNA molecules sequencing method performing [26]

Applying this technique allows us to obtain the count matrix directly. This matrix contains the amount of genes that correspond to the appropriate object and determines the gene expression value (level of gene activity). Then, we need to delete both the zero-expressed and low-expressed genes with the normalizing values of gene expression. The effectiveness of this step realization depends on both the boundary threshold index value that divides the genes into low-expressed and high-expressed ones and the selection of an appropriate normalizing method.

In [23], the solution of the problem regarding the formation of the gene expressions matrix using the results of the RNA molecules sequence test executed with the application of both Bioconductor package tools and data mining techniques has been dished.

However, we should note that the final effective decision of this problem can be achieved using current data mining and machine learning techniques successfully applied in various fields of intelligent data analysis nowadays [24-27].

4. Gene expression data statistical analysis and high-informative genes extraction

An analysis of various types of gene expression data allows concluding that in most of the cases initial dataset contains about 55000 genes but half of them have zero expression values for all of the investigated objects. After removing them, we have about 20000 genes. Removing low-expressed for all objects genes reduces their quantity to about 10000 ones. Thus, after the implementation of stage 2 (Fig. 1) the dataset contains about 10000 genes. This quantity is very large for gene regulatory network qualitative reconstruction or the creation of a qualitative disease diagnostic system.

The Cytoscape software tools [28] contain various techniques to allocate the co-expressed genes considering the patients' health state (investigated objects). In most cases, these techniques are based on cluster analysis algorithms. However, we should note that the selection of the following modelling genes is not an easy task, and its solution depends on both type of data and the goals of the determined problem.

The questions regarding the application of various biclustering techniques for allocation mutual correlated rows and columns have been presented in [29]. The authors compare different biclustering algorithms using both the synthetic data and gene expression profiles with an evaluation of the appropriate algorithm effectiveness in terms of quantitative biclustering quality criteria. The principal disadvantage of this technique is a large quantity of a few biclusters. Moreover, in most cases, the obtained biclusters do not contain all samples. This fact limited the successful implementation of this technique.

In [30], the questions regarding the creation of the hybrid model of gene expression profiling extraction using a statistical analysis technique, Shannon entropy, and fuzzy logic methods have been considered. The authors conducted a step-by-step manipulation of gene expression profiling removing with a calculation of clustering quality criteria which regarded both the density of the profiling concentration within clusters and density of the cluster centres distribution in the features space. The correlation distance was used as the proximity metric. They used as the quality measures the average of the profiling values for all samples and their variance were used as the statistical criteria. The Shannon entropy was taken as the third quality measure. The authors supposed that if the average and variance values are lesser and Shannon entropy are larger than the corresponding threshold then, this gene is deleted from the dataset as non-informative since it does not let us recognize the examined biological samples:

$$\mathrm{var} \le \mathrm{var}_{bound}\ ;\ aver \le aver_{bound}\ ;\ Sh\_entr \ge Sh\_entr_{bound} \qquad (1)$$

The threshold values of the respective criteria were determined using both clustering quality indexes and fuzzy logic model. The offered solution has allowed authors to allocate the most qualitative genes taking into account the resolvability of the studied patterns. These genes can be applied following for both the GRN reconstruction (reverse engineering task) and disease diagnostic system creation.

In [31], the authors have proposed a gene expression profiling selection technique using both data mining and machine learning techniques. A structural block-chart of the offered by the authors' general step-by-step process is introduced in Fig. 4.

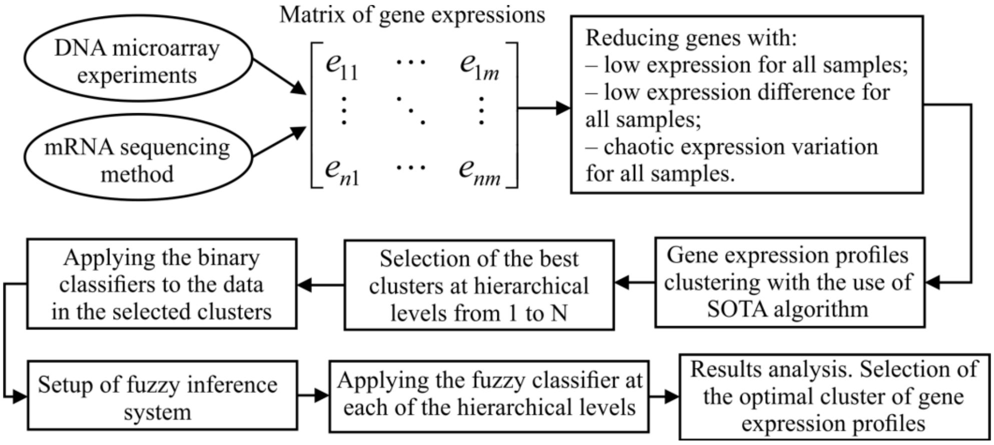
Matrix of gene expressions
DNA microarray experiments
mRNA sequencing method
e_{11} ⋯ e_{1m}
e_{n1} ⋯ e_{nm}
Reducing genes with:
– low expression for all samples;
– low expression difference for all samples;
– chaotic expression variation for all samples.
Gene expression profiles clustering with the use of SOTA algorithm
Selection of the best clusters at hierarchical levels from 1 to N
Applying the binary classifiers to the data in the selected clusters
Setup of fuzzy inference system
Applying the fuzzy classifier at each of the hierarchical levels
Results analysis. Selection of the optimal cluster of gene expression profiles

Figure 4: A step-by-step process of gene expression profiling selection [31]

The SOTA (Self Organized Tree Algorithm) was used as a clustering algorithm (data mining technique) [32,33]. The hierarchical clustering of the examined gene expression profiling at hierarchical levels from the first to tenth with an assessment of the clustering quality measure has been performed during the simulation procedure.

The optimal clusters matched to the minimum value of the quality measure was extracted at this step. Then, four classifiers (machine learning technique) were applied with the evaluation of the profiles classification quality using the ROC analysis technique. The conclusive solution regarding the selection of the clusters of gene expression profiling was taken applying a fuzzy inference system with the Mamdani inference algorithm, triangular and trapezoidal membership functions for input variables and only triangular membership functions for output measure. In the authors' opinion, the execution of the offered method can allow them to increase the efficiency of the procedure of co-

correlated gene expression profiling extraction. These profiles can be applied further for both GRN reconstruction (reverse engineering) and disease diagnostic systems creation.

5. Techniques of gene regulatory network reconstruction and modeling

In the common instance, a GRN is represented as a graph, where the nodes are control elements (genes, proteins, or metabolites), and the arcs measure the control interactions between the respective nodes (activation or deactivation procedures). Arcs can be directed, which indicates the nature of the interaction, and they can be weighted that indicates the strength of the appropriate interaction. Figure 5 shows the block-chart of existing patterns of GRN reconstruction and modelling in terms of the technique of their creation [34-37].

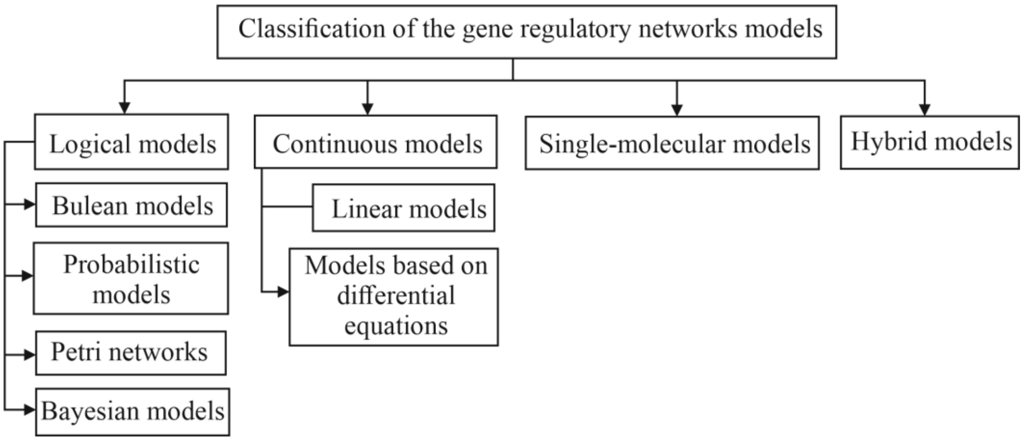

Figure 5: A block-chart of current patterns of GRN reconstruction and modelling

The analysis of Fig. 5 allows concluding that existing gene regulatory network models can be parted into logical models, which guess using the Boolean networks, probabilistic Boolean networks, and Bayesian networks, continuous models that guess using nonlinear and linear differential equations, models based on individual interconnected molecules, and hybrid models based on multilayer neuro-fuzzy networks. The selection of the suitable technique is certain by the amount and quality of data regarding gene expressions and relevant regulatory elements.

The boolean logical network presents itself as an active model of synchronical interactions between the appropriate network nodes. This class of model is one of the simplest models that allow presenting the particularities of actual gene networks of the examined objects [37,38]. Expression of gene values, in the case of Boolean network use, are replicated by logical values of 0 or 1. Zero value indicates this gene is not active and one value indicates the maximum degree of the gene activity. This indicates the fact that at the initial stage the data must be discretized. This fact is one of the principal shortcomings of this class of model since a large amount of useful information is lost during this model application. The advantages of this type of gene regulatory network are the simplicity of its use. Moreover, Boolean networks are comfortable for interpretation, they are suitable for the noised data processing because in this case, the sensitivity of the system is low.

A logical extension of a Boolean network is a probabilistic Boolean network [39,40]. The advisability of using this class of models can be valid by the fact that if the amount of information is incomplete or if the nature of the interaction between network nodes is insufficient understanding, the corresponding network nodes may have several regulatory functions. For this reason, these nodes have a certain degree of uncertainty. This fact can be quantified by the probability with which the node is expressed. At each step, the choice of regulatory function for each node of the network is determined by the corresponding probability, which depends on the values of a parental expression relative to this node of genes. We should note that this type of network partially takes into account the probability of different states of system nodes implementation, but the limitation of its applying as in the instance of the Boolean network depends on the need to determine the threshold of network gene expressions values.

As an alternative to Boolean and probabilistic networks in [41,42], the way of GRN reconstruction based on Petri networks was proposed. Modelling of the GRN using the Petri net is executed at the event level. In this case, the event means the transition of the system from one state to another one when the corresponding transitions are triggered. A transition is considered as open if, in each of its input positions, the number of markers is not so much as the number of arcs connecting the suitable node with this transition. Analysis of the results of network modelling allows us to determine the set of available states of the system and possible options for transition to the desired state. In the case of timing modelling, this means that the model based on the Petri net allows predicting the state of which genes and how it is necessary to change to achieve the desired overall state of gene regulatory network.

The last type of gene network logic models are models based on Bayesian networks [43]. The foundation for this class of network is the Bayesian rule, which involves, the expression values of the nodes can be represented using random variables within the range of the probability distribution. The Bayesian network can be defined as a directed acyclic graph, in which each of the nodes is represented a gene and each of the arcs is a probabilistic relation, which is a quantitative evaluation using appropriate conditional probabilities. The Bayesian network (BN) is founded on the Markov supposition that the value of parent nodes is not determined by the values of nodes that are not their descendants. The principal advantage of a Bayesian network is its ability to learn from existing data, and the relationships between variables can be linear, nonlinear, stochastic, combinatorial, and other types of interactions. They can allow us to model the gene network due to their ability to represent stochastic events and local processes of gene interaction in the presence of noise by determining the causal links between the respective nodes of the network.

The concept of modelling gene regulation using a system of equations assumes determining gene expression as a dependence function of the expression of other genes that interact with a given gene. A large number of works have been devoted to research in this subject area [44-46]. Gene regulatory network reconstruction and modelling by using systems of algebraic equations have some of the advantages. First, the equations of the system allow us to evaluate the regulatory processes in the network based on information about the expression of the corresponding genes. Moreover, a model based on a system of equations through positive and negative inverse relationships allows us to take into account the nature of the interactions, ie which genes act as activators and which as repressors. The main disadvantage of linear additive models is the circumstance that linear equations do not take into consideration the dynamic nonlinear aspect during genes regulation. In the instance of the high sensitivity of the pattern to variation of the value of gene expression, models based on differential nonlinear equations are more attractive. However, in the instance when we use a larger number of genes, debugging and explanation of the pattern is problematic.

Models founded on the interaction of single molecules are effective in the instance of both a small number of genes and the availability of sufficient information about the nature of the interaction of network elements. However, this fact limits its successful application. In the instance of the application of both the DNA microarray tests or RNA molecules sequencing experiments, the data contain tens of thousands of genes. pre-processing techniques can allow us to decrease the dimensionality of the attributes, but the number of genes used for gene network reconstruction in most instances limits the application of the patterns based on single molecules.

To compensate for the shortcomings inherent in discrete and continuous models, hybrid models have been proposed. These models take into account both the discrete and continuous aspects of gene regulatory network models. Thus, in [47,48], a hybrid model of a multilayer neuro-fuzzy recurrent network based on evolutionary learning algorithms was proposed. This model is focused on gene regulatory network reconstruction.

The advantages of this pattern include high computing power during information processing due to the application of a neural network. During the learning process, the model generates fuzzy rules based on existing data of gene expressions using evolutionary algorithms, which describe the real nature of gene interactions in the network. The accuracy of the network operation depends on both the choice of membership functions of the corresponding terms and the level of the range of gene expression variation. The main disadvantages of hybrid models are high levels of complexity and cost. Moreover, the sensitivity of these models significantly depends on their parameters, which increases the requirements for the process of model debugging.

6. Conclusions

In this review, we briefly described a step-by-step way of gene expression profiling processing got from experiments of both DNA microarray tests or RNA-molecules sequencing. The principal objective of the general process executing is gene regulatory network reconstruction and modelling or creation of patterns of diseases diagnostic. This problem is nowadays one of the current areas of bioinformatics. A qualitative reconstructed GRN can allow us to find out the nature of molecular elements interconnection what can lead to the following creation of both more effective complex diseases diagnostic systems and more acting medicines.

The analysis of the current research allows concluding that this problem has not a unique solution nowadays. The main intricacy of GRN reconstruction consists of the following: the experimental data which are used for the reconstruction procedure usually does not allow defining the network structure and pattern of genes interconnection in the network. Moreover, a large amount of genes intricates the explanation of the network nodes interconnections.

In beginning, we have introduced a general step-by-step way of gene expression profiling handling for GRN reconstruction (reverse engineering task), which assumes the offering of four stages. Then, we have briefly described each of the stages with the allocation of principal advantages and shortcomings.

The following perspectives of the authors' research are the development of gene expression profiling data processing methods for purpose of extraction of most active genes considering the type of the disease and the creation of more acting techniques of GRN reconstruction and modelling based on the application of the extracted genes.

7. References

[1] Yang, L., Yang, Y., Meng, M., et al.: Identification of prognosis-related genes in the cervical cancer immune microenvironment, Gene, 766, art. no. 145119 (2021) doi: 10.1016/j.gene.2020.145119

[2] Taniguchi-Ponciano, K., Vadillo, E., Mayani, H., et al.: Increased expression of hypoxia-induced factor 1α mRNA and its related genes in myeloid blood cells from critically ill COVID-19 patients Annals of Medicine, 53(1), pp. 197-207 (2021) doi: 10.1080/07853890.2020.1858234

[3] Yu, X., Abbas-Aghababazadeh, F., Chen, Y.A., et al.: Statistical and bioinformatics analysis of data from bulk and single-cell RNA sequencing experiments. Methods in Molecular Biology, 2194, pp. 143-175 (2021) doi: 10.1007/978-1-0716-0849-4_9

[4] Diroma, M.A., Ciaccia, L., et al.: Bioinformatics resources for RNA editing Methods in Molecular Biology, 2181, pp. 177-191 (2021) doi: 10.1007/978-1-0716-0787-9_11

[5] Cha, Y.J., Park, S.M., et al.: Microstructure arrays of DNA using topographic control. Nat. Commun., 10(1), art. no. 2512 (2019) doi: 10.1038/s41467-019-10540-2

[6] Nagashima, M., Miwa, N., Hirasawa, H., et al.: Genome-wide DNA methylation analysis in obese women predicts an epigenetic signature for future endometrial cancer. Sc.. Rep., 9(1) (2019), art. no. 6469 doi: 10.1038/s41598-019-42840-4

[7] Depledge, D.P., Srinivas, K.P., Sadaoka, T., et al.: Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat. Communic., 10(1), art. no. 754 (2019) doi: 10.1038/s41467-019-08734-9

[8] Lian, B., Hu, X., Shao, Z.-M.: Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. Sc. Rep., 9(1), art. no. 6032 (2019) doi: 10.1038/s41598-019-42184-z

[9] Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.. Bioinf., 19 (2), pp. 185-193 (2003) doi: 10.1093/bioinformatics/19.2.185

[10] Affymetrix. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA, pp. 1-27 (2002)

[11] Irizarry, R.A., Hobbs, B., Collin, F., et al.: Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. Sel. Works of Terry Speed, pp. 601-616 (2012) doi: 10.1007/978-1-4614-1347-9_15

[12] Buermans, H.P.J., Ariyurek, Y., van Ommen, G., et al.: New methods for next generation sequencing based microRNA expression profiling. BMC Gen., 11(1), art. no. 716 (2010) doi: 10.1186/1471-2164-11-716

[13] Hackenberg, M., Sturm, M., Langenberger, D., et al.: miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. Nucl. Ac. Res., 37 (SUPPL. 2), pp. W68-W76 (2009) doi: 10.1093/nar/gkp347

[14] Farazi, T.A., Brown, M., Morozov, P., et al.: Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. Meth., 58 (2), pp. 171-187 (2012) doi: 10.1016/j.ymeth.2012.07.020

[15] Hackenberg, M., Rodríguez-Ezpeleta, N., Aransay, A.M.: MiRanalyzer: An update on the [5] detection and analysis of microRNAs in high-throughput sequencing experiments. Nucl. Ac. Res., 39 (SUPPL. 2), pp. W132-W138 (2011) doi: 10.1093/nar/gkr247

[16] Huang, P.-J., Liu, Y.-C., Lee, C.-C., et al.: DSAP: Deep-sequencing small RNA analysis pipeline. Nucl. Ac. Res., 38 (SUPPL. 2), art. no. gkq392, pp. W385-W391 (2010) doi: 10.1093/nar/gkq392

[17] Morin, R.D., O'Connor, M.D., Griffith, M., et al.: Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Gen. Res., 18 (4), pp. 610-621 (2008) doi: 10.1101/gr.7179508

[18] ArrayExpress. El. Resource: https://www.ebi.ac.uk/arrayexpress/

[19] El. Resource: https://microbenotes.com/dna-microarray/

[20] Babichev, S., Durnyak, B., Zhydetskyy, V., Pikh, I., Senkivskyy, V.: Techniques of DNA microarray data pre-processing based on the complex use of Bioconductor tools and Shannon entropy. CEUR Workshop Proceedings, 2353, pp. 365-377 (2019)

[21] Hausser, J., Strimmer, K.: Entropy inference and the James-stein estimator, with application to nonlinear gene association networks. J. of Mach. Learn. Res., 10, pp. 1469-1484 (2009)

[22] El. Resources: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/rna-seq

[23] Babichev, S., Durnyak, B., Senkivskyy, V., et al.: Exploratory analysis of neuroblastoma data genes expressions based on Bioconductor package tools. CEUR Workshop Proceedings, 2488, pp. 268-279 (2019)

[24] Tkachenko, R., Izonin, I., Kryvinska, N., et. al.: An approach towards increasing prediction accuracy for the recovery of missing iot data based on the grnn-sgtm ensemble. Sensors (Switzerland), 20 (9), art. no. 2625, (2020) doi: 10.3390/s20092625

[25] Izonin, I., Tkachenko, R., Verhun, V., et al., An approach towards missing data management using improved GRNN-SGTM ensemble method, Intern. J. Engineering Science and Technology, 2020. https://doi.org/10.1016/j.jestch.2020.10.005 (in press)

[26] Rzheuskyi, A., Kutyuk, O., Vysotska, V., et al.: The Architecture of Distant Competencies Analyzing System for IT Recruitment. IEEE 2019 14th Intern. Sc. and Techn. Conf. on Comp. Sc. and Informat. Technol., CSIT 2019 - Proceedings, 3, art. no. 8929762, pp. 254-261 (2019) doi: 10.1109/STC-CSIT.2019.8929762

[27] Lytvyn, V., Salo, T., Vysotska, V., et al.: Identifying Textual Content Based on Thematic Analysis of Similar Texts in Big Data. IEEE 2019 14th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2019 - Proceedings, 2, art. no. 8929808, pp. 84-91 (2019) doi: 10.1109/STC-CSIT.2019.8929808

[28] CytoScape Homepage. El. Resource: http://apps.cytoscape.org

[29] Babichev, S., Osypenko, V., Lytvynenko, V., et al.: Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. 2018 IEEE 38th Intern. Conf. on Electr. and Nanotechn., ELNANO 2018 - Proceedings, art. no. 8477439, pp. 298-304. (2018) doi: 10.1109/ELNANO.2018.8477439

[30] Babichev, S., Barilla, J., Fišer, J., Škvor, J.: A hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criteria. Proc. of the 11th Conf. of the Europ. Soc. for Fuzzy Log. and Technol., EUSFLAT 2019, pp. 128-133 (2020)

[31] Babichev, S., Škvor, J.: Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. Diagnostics, 10 (8), art. no. 584 (2020) doi: 10.3390/diagnostics10080584

[32] Dorazo, J., Carazo, J.M.: Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J. of Molec. Evolut., 44(2), pp. 226–260 (1997) doi: 10.1007/PL00006139

[33] Fritzke, B.: Growing cell structures a self-organizing network for unsupervised and supervised learning. Neural Networks, 7(9), pp. 1441–1461 (1994) doi: 10.1016/0893-6080(94)90091-4

[34] De Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol., №9(1). pp. 67–103 (2002)

[35] Van Someren, E. P., Wessels, L. F., Backer, E., Reinders, M. J.: Genetic network modeling. Pharmacogenomics, Vol. 3. pp. 507–525 (2002)

[36] Schlitt, T., Brazma, A.: Current approaches to gene regulatory network Modeling. BMC Bioinf., № 8(6). pp. 1–22 (2007)

[37] Nedumparambathmarath, V., Chakrabarti, S. K., et al.: Modeling of gene regulatory networks: A review. J. Biomed. Sc. and Engin., Vol. 6. P. 223–231 (2013)

[38] Faure, A., Naldi, A., et al.: Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics, Vol. 22. pp.124–131 (2006)

[39] Shmulevich, I., Dougherty, E.R., et al.: Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinf., Vol.18. pp. 261–274 (2002)

[40] Shmulevich, I., Gluhovsky, I., Hashimoto, R. F., Dougherty, E. R., Zhan, W.: Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks. Comparative and Functional Genomics, Vol. 4. pp. 601–608 2003

[41] Reddy, V. N., Liebman, M. N., et al.: Qualitative analysis of bio-chemical reaction systems. Computational Biology and Medical Informatics, Vol. 26. pp. 9–24 (1996)

[42] Remy, E., Mendoza, L., Thieffry, D., Chaouiya, C.: From logical regulatory graphs to standard Petri nets: Dynamical roles and functionality of feedback circuits. Lecture notes in Computer Science, Vol. 4230. pp. 56–72 (2006)

[43] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyse expression data. J. of Computat. Biol., Vol. 7. pp. 601–620 (2000)

[44] Chen, T., He, H. L., Church, G. M.: Modeling gene expression with differential equations. Proceedings of the Pacific Symposium on Biocomputing, Vol. 4. pp. 29–40 (1999)

[45] Van Someren, E. P., Vaes, B. L. T., Steegenga, W. T. et al.: Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics, Vol. 22(4), pp. 477–484 (2006)

[46] Sakamoto, E., Iba, H.: Inferring a system of differential equations for a gene regulatory network by using genetic programming. Proc. of the Congress on Evolutionary Computation, pp. 720–726 (2001)

[47] Ioannis, A. M., Andrei, D., Dimitris, T.: Gene regulatory networks Modeling using a dynamic evolutionary hybrid. BMC Bioinformatics, Vol. 11, pp. 1–17 (2010)

[48] Du, P., Gong, J., et al.: Modeling gene expression networks using fuzzy logic. IEEE Trans. on Syst., Man and Cybern., Vol. 35, pp. 1351–1359 (2005)