UDC 004.048+004.94+577
DOI: 10.15587/1729-4061.2018.123634

DEVELOPMENT OF A TECHNIQUE FOR THE RECONSTRUCTION AND VALIDATION OF GENE NETWORK MODELS BASED ON GENE EXPRESSION PROFILES

S. Babichev
Associate professor
Department of Informatics
Jan Evangelista Purkyně University
Pasteurova 1, Ústí nad Labem, Czech Republi, 40096
PhD, Associate Professor
Department of Informatics and Computer Science
Kherson National Technical University
Beryslavske highway, 24, Kherson, Ukraine, 73008
E-mail: sergii.babichev@ujep.cz
M. Korobchynskyi
Doctor of Technical Sciences, Professor, Senior Researcher
Military-Diplomatic Academy named after Eugene Bereznyak***
Melnykova str., 81, Kyiv, Ukraine, 04050
E-mail: maks_kor@ukr.net
O. Lahodynskyi
Doctor of Pedagogical Sciences,
Associate Professor, Head of Department**
E-mail: berezan2016@meta.ua
O. Korchomnyi
PhD, Associate Professor*
E-mail: alxkor53@gmail.com
V. Basanets
PhD, Associte Professor**
E-mail: vlad_bas@rambler.ru
V. Borynskyi
PhD, Associte Professor**
E-mail: vovik79@ukr.net
*Department No. 5***
Department No. 6*
***Military-Diplomatic Academy named after Eugene Bereznyak
Melnykova str., 81, Kyiv, Ukraine, 04050

Розроблено технологію реконструкції та валідації моделей генних мереж на основі профілів експресій генів. Представлені дослідження по оптимізації топології генної мережі на основі комплексного застосування топологічних параметрів мережі та функції бажаності Харрінгтона. Запропонована технологія валідації моделі генної мережі на основі ROC-аналізу, реалізація якої передбачає порівняльний аналіз характеру зв'язків між відповідними генами у реконструйованих мережах

Ключові слова: генна мережа, топологічні параметри, індекс бажаності Харрінгтона, експресія генів, коефіцієнт трешолдингу

Разработана технология реконструкции и валидации моделей генных сетей на основе профилей экспрессии генов. Представленны исследования по оптимизации топологии генной сети на основе комплексного использования топологических параметров сети и функции желательности Харрингтона. Предложена технология валидации модели генной сети на основе ROC-анализа, реализация которой предусматривает сравнительный анализ характера связей между соответствующими генами в реконструированых генных сетях

Ключевые слова: генная сеть, топологические параметры, индекс желательности Харрингтона, экспрессия генов, коэффициент трешолдинга

1. Introduction

Modern information processing systems in most cases are based on the use of analogies of functioning of biological mechanisms and processes occurring in living organisms. These processes include the functioning of a natural neural network, immune processes, a gene network, etc. Special feature of such systems is the decentralized parallel information processing, high level of complexity, learning ability, capabilities to recognize information and form decisions. The creation of artificial models of modern biological systems can be based on the systems approach, which implies the integrated application of methods from molecular biology, mathematics, computer science, the laws of physics. The implementation of a given approach creates conditions for understanding which factors determine the character of functioning of a biological system in order to correct the process.

Reconstruction and simulation of a gene regulatory network (GRN) underlies studies and analysis of the character of gene interaction and the effects of these interactions on

the functionality of a biological organism. The complexity of a GRN reconstruction process is predetermined by the fact that experimental data, which are used for the reconstruction of the network, do not typically make it possible to unambiguously define the structure of a network and the character of relationship between the nodes. In addition, a large number of genes that determine the structure and size of the network complicate the process of interpretation of the results obtained. Therefore, there is a need for research into quantitative estimation of network topology and the character of relations between the elements using, as experimental data, the profiles of gene expressions derived from DNA-microchip experiments.

There are databases of biological gene regulatory networks of different organisms [1] that make it possible to visualize and explore a network topology of the appropriate biological object under study. One of modern methods employed for the reconstruction of GRN is the identification of a network by comparing it with a known network of the relevant biological organism. In this case, the criteria for comparison are the network topology and the existence of appropriate genes in a given network compared to a reference network. Another important stage when reconstructing a network is the validation process based on quantitative criteria for assessing the network quality. A structural block diagram of a general process of information processing aimed at the reconstruction and validation of GRN is shown in Fig. 1.

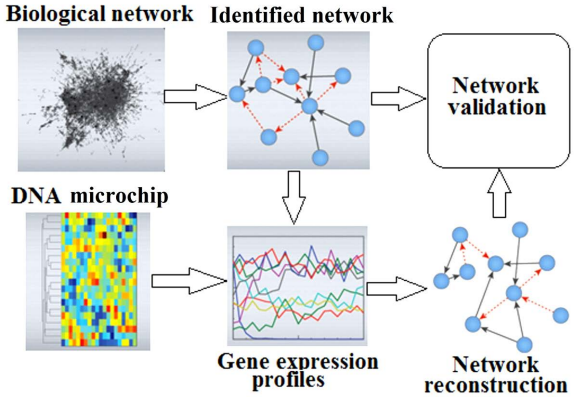

Fig. 1. Structural block diagram of a general process of reconstruction and validation of the model of a gene regulatory network

The process of GRN reconstruction and validation implies a comparative analysis of parameters of the reconstructed and reference networks with the purpose of determining an optimal network topology and the character of interrelations between relevant nodes. The research performed in order to create a technology for GRN reconstruction based on statistical methods for estimating the strength and character of interactions between genes makes it possible to optimize parameters of the algorithm for reconstructing a gene network. Accurately reconstructed GRN makes it possible to explore the character of development of a biological organism at the gene level, which creates preconditions for early diagnosis and adjustment of the development of different types of genetic diseases. This fact demonstrates the relevance of research topic presented here.

2. Literature review and problem statement

Modern biological systems of living organisms are a complex dynamic network of interacting elements with different purposes whose state can change under the influence of external conditions [2]. Reconstruction and modeling of gene networks are rather complicated tasks that do not have an unambiguous solution at present. The first studies on the reconstruction of biological networks based on experimental data were published at the end of the 90s of the last century [3–6]. Those papers proposed several approaches to a given type of modeling. Studies [7–9] reviewed several methods related to the reconstruction and modeling of GRN models based on data about gene expressions. The authors considered in detail stages in the process of reconstructing gene networks and performed a comparative analysis of different methods with outlined advantages and shortcomings of the respective method.

The basic idea of GRN reconstruction is to use experimental data on gene expressions in order to obtain the models through estimation and analysis of the bonds between molecular objects. It should be noted, however, that a given process is very complicated due to the fact that this problem is of a combinatorial character, on the one hand, and that experimental data in many cases are incomplete and inaccurate, on the other hand. In addition, the existence of a large number of parameters, variables and constraints, necessitates the application of numerical and computational methods.

Modern technologies for obtaining data on gene expressions tend to cover the maximum number of variables in the system [10]. For example, DNA microchip technology makes it possible to measure the expression of tens of thousands of genes simultaneously, that is, each examined object is characterized by the numerical vector of gene expressions with a length of tens of thousands of units. Such a large number of parameters is predetermined by a variety of processes occurring in a biological system. At present, there is sufficient information about the properties of gene regulatory networks in natural biological systems. In papers [11, 12], authors formulated the rules for GRN reconstruction and modeling, which make it possible to significantly limit the dimensionality of search space for the optimal network. The sparsity property is the most common and important feature of GRN. This property means that the topology of GRN is sparse, meaning that each gene has a small number of regulatory inputs [13]. It should be noted, however, that there are a small number of genes (master-genes) that are capable of controlling hundreds of other genes. A given property is used to limit the search space of the optimal solution by limiting the number of regulatory bonds. Papers [14, 15] show that the frequency distribution of the number of regulatory inputs of nodes at a gene network of biological systems is often governed by the law of Pareto distribution. This means that in the case of a non-scalable network most genes are loosely linked, but there are several nodes with a high number of links – nodes-concentrators. These nodes correspond to genes that perform most of the overall regulation of other nodes in the network. The existence of a concentrator leads to the localization of the network, because all the nodes in the network are connected to concentrators via short bonds, the quantity of which is limited. In addition, concentrators improve stability of the model against external influences and various kinds of fluc-

tuations, since they bind a network and do not provide for the possibility to split the network into separate fragments. The next property of GRN, which must be considered when reconstructing a network, is modularity. This property means that genes in the network cannot be regarded as independent elements. In a general case, genes can be divided into functional, perform the function of control over other genes (concentrators), and genes that function in concert, performing a joint function. It is obvious that in this case genes can be grouped in modules or clusters depending on the functional similarity in the profiles of expressions. Technology of a bicluster analysis [16], which is widely used now for grouping the genes and objects, does not resolve the task on grouping the profiles of gene expressions. As shown by the authors, the use of algorithms for a bicluster analysis allows obtaining clusters of mutually-correlated genes and objects, but there is a problem of the choice of the number of biclusters and the level of detail of this process. Paper [17] reports results of the study into determining the optimal affinity function in order to estimate the degree of closeness in the profiles of gene expressions. The effective criteria are specified for estimating the clustering quality of profiles of gene expressions that create preconditions for enhancing objectivity when grouping high-dimensionality complex data. In [18], authors used the example of model data to run a comparative analysis of internal and external quality criteria when clustering the profiles of gene expressions, which made it possible to propose an integrated multiplicative criterion of quality for assessing the grouping of complex objects. The results of practical implementation of the study conducted are given in [19, 20]. The authors developed the inductive technology for objective clustering of the profiles of gene expressions [19] and implemented this technology in practice applying the density algorithm DBSCAN algorithm and the self-organizing tree algorithm SOTA [20].

The result of simulation is the proposed model of a cluster-bicluster analysis, implementation of which makes it possible to increase the amount of useful information for the further reconstruction of a gene network.

Based on analysis of the scientific literature, we can conclude that at present there is no any effective technology for the reconstruction of a gene regulatory network that is capable, with a high degree of probability, of predicting the character of further development of a biological organism at the gene level. GRN topology is defined by the type of the appropriate algorithm and its parameters. Therefore, the development of effective technology for determining optimal parameters of a GRN reconstruction algorithm, as well as the validation method for derived models of gene networks, is one of the promising tasks in modern bioinformatics, solving which would improve the effectiveness of diagnosis and treatment of complex genetic diseases.

3. The aim and objectives of the study

The aim of present study is to develop a technology for the reconstruction and validation of a gene regulatory network based on statistical methods of analysis of the character of interrelations between respective genes.

To achieve the set aim, the following tasks have been solved:

– to develop a technology for the reconstruction of gene regulatory networks based on comprehensive use of network topology parameters and the generalized desirability indicator by Harrington;

– to develop a validation technology for the models of gene networks based on ROC analysis whose implementation implies a comparative analysis of the character of relations between respective genes in the network based on the entire totality of genes and gene networks based on the obtained biclusters;

– to model the processes of reconstruction and validation of models of gene networks using gene expression profiles.

4. Estimation parameters for the gene regulatory network topology

The process of gene regulatory network reconstruction based on gene expression profiles implies a possibility to create different topologies of networks that differ from each other by the number of nodes, the number of arcs that connect respective nodes, and character of bonds between the nodes of the network. As a result, it is necessary to quantify a network topology, that is, identify parameters that make it possible to reasonably select optimal network topology for the respective biological object. An example of the biological gene regulatory network topology is shown in Fig. 2 [21].

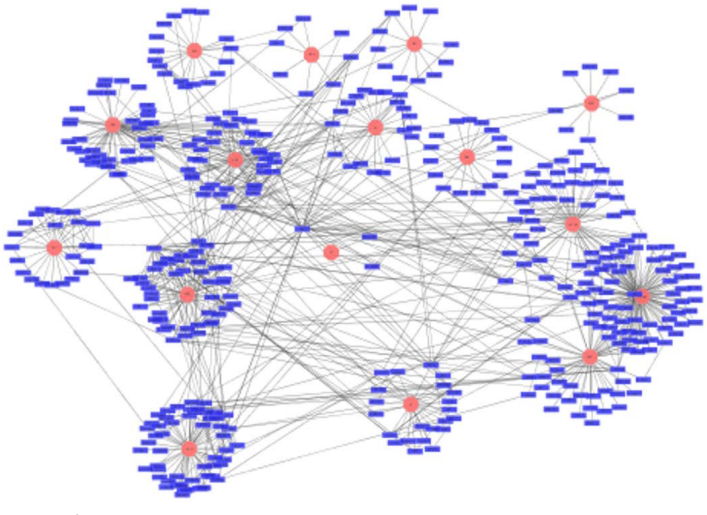

Fig. 2. Example of a biological gene regulatory network topology

Analysis of the structure and topology of GRN allows us to conclude that it is a directed or a non-directed graph whose arcs can be weighted (in the presence of weight that determines the strength of the connection), or weightless. Therefore, to identify the parameters that determine a network topology, we can apply a graph theory. Classification of basic topological characteristics of GRN is shown in Fig. 3 [22].

The number of nodes in a network determines the total number of interconnected genes and may differ from the total number of genes in the network. A gene-specific parameter that determines the strength of the relationship with neighboring genes may be less than the threshold value set in the process of forming the network. In this case, this gene

is separated and removed from further research. In addition, the network may contain few genes, which are linked, but these genes do not have connections between genes that make up the basic network. These genes can also be removed from further research.

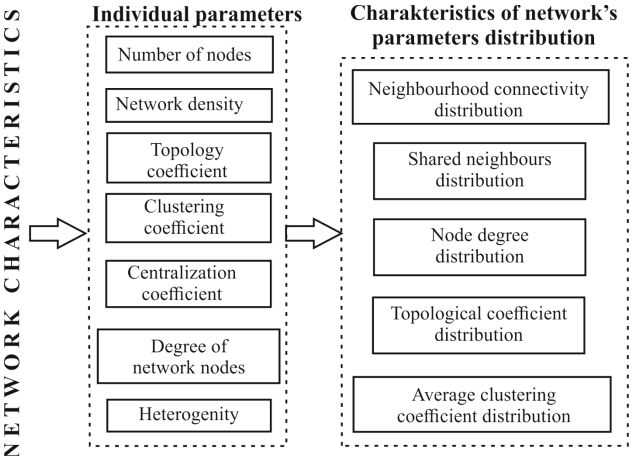

Fig. 3. Classification of characteristics to assess the topology of a gene regulatory network

The degree of a node in the network, or its *connectivity,* is the total weight of links (arcs), connecting a given node with neighboring nodes:

$$k_i = \sum_{j=1, j\neq i}^{n_i} \omega_{ij}, \tag{1}$$

where n_i is the number of nodes of the i-th gene, ω_{ij} is the weight of the arc that connects neighboring genes i and j.

The average degree, or *mean connectivity,* of the network average is defined as the mean value of degrees of all nodes in the network:

$$k_{\text{cep}} = \frac{1}{n}\sum_{i=1}^{n} k_i. \tag{2}$$

Maximum degree determines the maximum value of elements of the connectivity vector of all nodes:

$$k_{\max} = \max\left(k_1, k_2, ..., k_n\right). \tag{3}$$

The high value of connectivity of the network indicates a high level of complexity because all the nodes of the network have a large number of relations with neighbors. This fact complicates the interpretation of the resulting network. Obviously, the value of a given parameter are optimal in the case of the presence of a minimum under condition of the unalterable number of genes contained in the nodes of the network.

The density of the network is defined as the ratio of the number of weighted connections between the nodes of the network to the largest possible number of connections between the nodes in a given network:

$$D = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \omega_{ij}}{0{,}5\times n\left(n-1\right)}, \tag{4}$$

where ω_{ij} is the weight factor between nodes i and j, and n is the total number of nodes in the network. Value of a weight coefficient varies from 0 to 1. If this parameter is null, then there is no connection between the relevant nodes. Value of a network density also varies from 0 to 1. If D=0, there is no connection between genes; if D=1, we have a fully connected network. It is obvious that a decrease in parameter D at a constant number of genes in the network indicates a decrease in the number of links in the network, thus increasing the time of the transfer of information, and the process of interpretation of the network is simplified.

A node clustering coefficient determines the probability that the next neighbors of a given node are linked directly. *A network clustering coefficient* is defined as the average of clustering coefficients of all nodes:

$$C = \frac{1}{n}\sum_{i=1}^{n} \frac{e_i}{0{,}5\times k_i\left(k_i-1\right)}, \tag{5}$$

where n is the number of genes in the network, e_i is the number of actual connections between node i and neighboring nodes, k_i determines the number of neighbors of gene i, this gene including, which can form a complete cluster. This parameter is a quantitative measure for the fully connected network. If its value is equal to unity, the network is fully connected; with a zero value, the network has no links to the neighbors of network's genes.

A network centralization coefficient determines the degree of proximity to the star topology. This coefficient is calculated from formula:

$$\text{Centr} = \frac{n}{n-2}\left(\frac{k_{\max}}{n-1} - D\right). \tag{6}$$

If a network topology takes the shape of a grid where all nodes are similarly connected, the value of a given parameter is zero. A higher value of centralization parameter corresponds to a higher degree of similarity of the network to a star-shaped topology.

Heterogeneity of the network determines the degree of heterogeneity of network topology and is expressed through the variance and mean of the average degree of the nodes by formula:

$$G = \frac{\sqrt{\text{var}\left(k_{\text{cep}}\right)}}{\text{mean}\left(k_{\text{cep}}\right)}. \tag{7}$$

A homogeneous network has zero heterogeneity, the growth of the value of this parameter indicates a greater difference between the values of degrees of the network nodes.

Coefficient of topology of node n, which has k_n neighbors, is determined from formula:

$$T_n = \frac{\frac{1}{m}\sum_{i=1}^{m} L\left(n,i\right)}{k_n}, \tag{8}$$

where m is the number of nodes that have common neighbors with gene n, $L(n, i)$ determines the number of neighbors of gene i that have at least one neighbor with gene n.

Values of individual parameters for a comprehensive assessment of the topology of a gene regulatory network, shown in Fig. 2, are given in Table 1.

The indicated parameters make it possible to make a preliminary assessment of the GRN model topology. At a constant number of nodes, lower values of density and clustering

of the network and a larger heterogeneity value testifies to the higher quality of network topology. A higher value of centralization coefficient indicates the degree of proximity of network topology to a star-shaped structure.

Table 1

Estimation parameters for the topology of a biological gene regulatory network

Number of nodes	Network density	Clustering coefficient	Centralization coefficient	Heterogeneity
471	0.008	0.213	0.139	1.967

Analysis of values of topological parameters of a gene network makes it possible to define steps for the formation of a network topology based on gene expressions that make up the backbone of the network. On the one hand, the network should contain the maximum number of examined genes (after the processes of filtration, reduction, clustering and biclustering). On the other hand, network density and clustering coefficient should be minimal, and the coefficients of heterogeneity and centralization – maximum. These rules will underlie the creation of technology for the reconstruction of a gene regulatory network based on gene expression profiles.

5. Reconstruction of a gene regulatory network based on correlation analysis

The process of GRN reconstruction based on correlation analysis implies the calculation of coefficients of pair correlation between the examined gene expression profiles. Since in the case of analysis of the matrix of gene expressions the vectors of profiles are the sequences of rational numbers, it is appropriate to use the Pearson method for calculating a pair correlation between respective profiles:

$$r(X_a, X_b) = \frac{\sum_{i=1}^{m} (x_{ai} - \bar{x}_a)(x_{bi} - \bar{x}_b)}{\sqrt{\sum_{i=1}^{m} (x_{ai} - \bar{x}_a)^2} \cdot \sqrt{\sum_{i=1}^{m} (x_{bi} - \bar{x}_b)^2}}, \quad (9)$$

where X_a, X_b are the vectors of the examined gene expressions profiles, m is the number of attributes in the respective vectors, $\bar{x}_a$, $\bar{x}_b$ represent the average values of profiles of X_a, X_b, respectively. The coefficient of pair correlation in the case of its significance represents the strength of the relationship between the corresponding nodes of the network. When using a full matrix of coefficients of pair correlation, a gene network is fully connected, since there is connection between all the nodes of a given network. The weight of the arc is equal to the coefficient of correlation between a pair of gene expression profiles whose relations are assessed. Network topology in this case is determined by the value of threshold coefficient τ that defines the threshold value of the existence of a relationship between a pair of genes in the network. A weight factor of the arc that connects the corresponding genes is defined as follows:

$$\omega(X_a, X_b) = \begin{cases} 0, \text{ if } r(X_a, X_b) < \tau; \\ r(X_a, X_b), \text{ if } r(X_a, X_b) \geq \tau. \end{cases} \quad (10)$$

Simulation of the process of a gene network reconstruction based on gene expression profiles was performed in the programming environment *CytoScape* [22] using the gene expression profiles data *moe430a* from the database *ArrayExpress* [23]. The data were acquired from DNA-microchip experiments and contained information on the gene expression of mesenchymal cells of two types: nerve crest and mesoderm. The matrix of original data consisted of 147 lines, or genes, and 20 columns, or the examined objects. A block diagram of the simulation process of GRN reconstruction modeling based on the correlation output algorithm is shown in Fig. 4. Implementation of a given algorithm implies the following steps:

1. Formation of the input data in the form of a matrix where lines are the genes that represent nodes of a gene network, and columns are the conditions for running an appropriate experiment to determine the expressions of genes.

2. Assigning the interval and step for a change in the value of the threshold coefficient; initializing the initial value of threshold coefficient $\tau = \tau_{min}$.

3. Reconstruction of GRN whose topology matches the assigned value of threshold coefficient.

4. Calculation of topological parameters of the obtained GRN in line with formulae (1)–(8).

5. If the value of the threshold coefficient is less than the maximum, we increase this value by $d\tau$ (a step of change in the threshold coefficient) and proceed to step 3 of a given procedure. In the opposite case, we construct diagrams of dependence of the obtained topological parameters on the value of threshold coefficient and diagrams of distributed topological parameters for each value of the topological coefficient.

6. Analysis of the obtained results, determining the value of threshold coefficient that matches the optimal GRN topology.

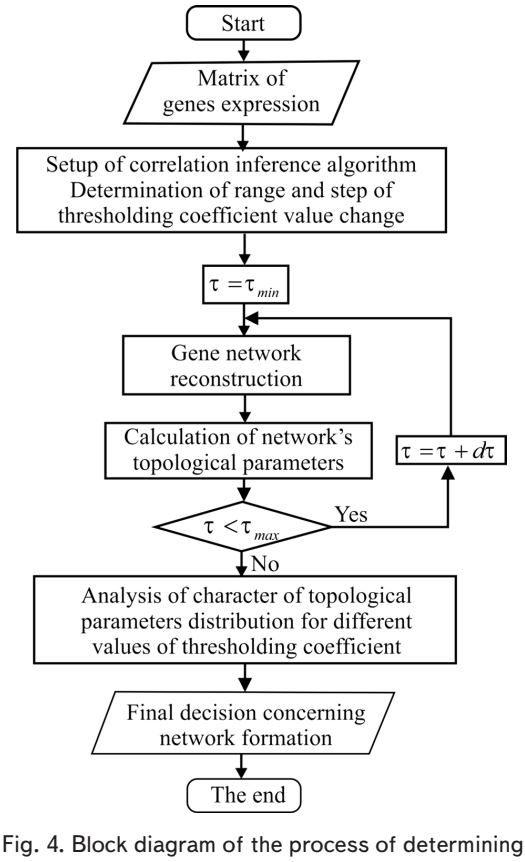

Fig. 4. Block diagram of the process of determining the optimal value of threshold coefficient when using a correlation output algorithm

Fig. 5 shows results of the algorithm execution in the form of diagrams of dependence of individual topological parameters on the threshold coefficient. Value of the threshold coefficient changed from 0.3 to 0.7 with a step of 0.05. An analysis of the acquired diagrams reveals that in the interval of change in the values of threshold coefficient from 0.3 to 0.45 the number of genes in the network does not change. In this case, the values of coefficients of centralization and heterogeneity grow while those of clustering and density coefficients decrease. This indicates improvement in the network topology by reducing the number of connections between its nodes at a constant number of genes. When the value of threshold coefficient is 0.5, the number of genes is reduced from 147 to 146 while the centralization coefficient reaches its maximum. Upon further increase in the value of threshold coefficient, the number of genes and the value of centralization coefficient start to decline sharply. This fact testifies to the deterioration of the network topological structure.

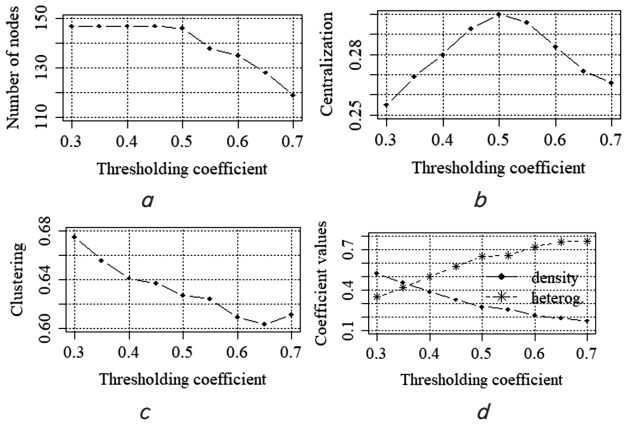

Fig. 5. Diagrams of dependence of individual parameters for the estimation of a gene network topology on the values of threshold coefficient: *a* — number of genes; *b* — centralization coefficient; *c* — clustering coefficient; *d* — density and heterogeneity of the network

The research conducted allowed us to define a narrower interval for a change in the value of threshold coefficient for determining the optimal topology of a gene network.

Fig. 6 shows diagrams of change in the individual topological parameters at a change in the value of threshold coefficient from 0.45 to 0.55 with a step of 0.01. In this case, in the presence of several genes, linked together but separated from the main network, we separated a network of genes that have the largest number of interconnected nodes. Subnetwork with multiple nodes was removed from the network.

Fig. 6 shows that the values of threshold coefficient that defines the structure of a gene network is determined by four topological parameters: coefficients of clustering, centralization, the network homogeneity, and nodes density. It should be noted that the optimal network structure corresponds to the minimum values of nodes density and clustering coefficient, and to the maximum values of centralization coefficient and heterogeneity coefficient. In order to make a final decision on the choice of a network structure, it is proposed to employ a comprehensive criterion based on the Harrington desirability function [23]. The application of a given function implies the conversion of scales for topological parameters into a linear scale of dimensionless indicator Y whose value varies from −2 to 5. Private desirables for each value of indicator Y are calculated from formula:

$$d = \exp\left(-\exp\left(-Y\right)\right). \tag{11}$$

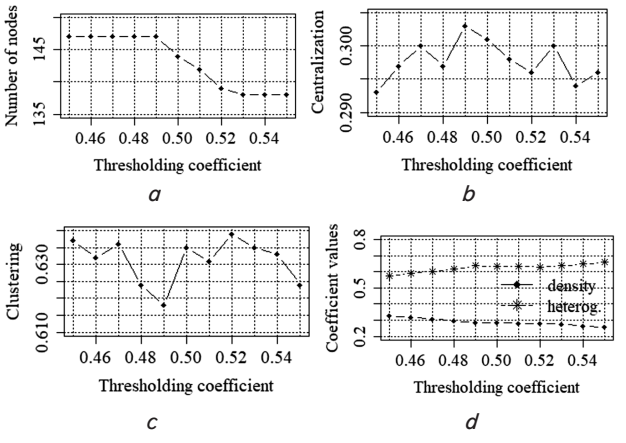

Fig. 6. Diagrams of change in individual parameters for the estimation of a gene network topology in the interval of change in threshold coefficient from 0.45 to 0.55 with a step of 0.01: *a* — number of genes; *b* — centralization coefficient; *c* — clustering coefficient; *d* — density and heterogeneity of the network

Desirability scale has an interval from 0 to 1. Value d=0 indicates an absolutely impossible topology in terms of a given criterion; d=1 is the best topology. The choice of points 0.63 and 0.37 on the scale of desirability is due to the convenience of calculations: 0.63=1–1/e, and 0.37=1/e. The value of 0.37 typically corresponds to the limit of permissible values. Fig. 7 shows chart of the Harrington desirability function constructed according to formula (11).

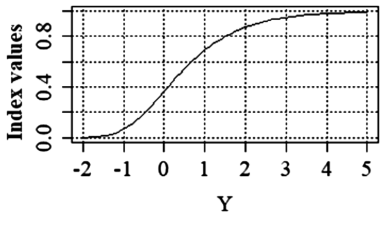

Fig. 7. Harrington desirability function

Practical implementation of the calculation algorithm for a comprehensive criterion based on the network topological parameters implies the following steps:

1. Transformation of scales of network topological parameters to the scale of dimensionless indicator Y in accordance with a system of linear equations:

$$\begin{cases} Y_{dens} = a_1 - b_1 \cdot ds; \\ Y_{clust} = a_2 - b_2 \cdot cl; \\ Y_{centr} = a_3 + b_3 \cdot cr; \\ Y_{hetr} = a_4 + b_4 \cdot hr, \end{cases} \tag{12}$$

where *ds, cl, cr* and *hr* are the values of parameters for the density of nodes, clustering, centralization and heterogeneity, which were calculated applying the corresponding value of threshold coefficient; a and b are parameters that

are determined empirically for each criterion based on its boundary values. In the case of density of nodes and clustering coefficient, a system of equations for determining parameters a and b takes the form:

$$Y_{\max} = a - b \cdot X_{\min},$$

$$Y_{\min} = a - b \cdot X_{\max}. \quad (13)$$

When calculating a generalized indicator based on the coefficients of centralization and heterogeneity, the system of equations (13) takes the form:

$$Y_{\max} = a + b \cdot X_{\max},$$

$$Y_{\min} = a + b \cdot X_{\min}, \quad (14)$$

where $Y_{\min}=-2$, $Y_{\max}=5$ (boundary values for the scale of the generalized indicator, which match the values of Harrington desirability function of 0 and 1, respectively), X_{min} and $X_{\max}$ are the minimum and maximum value of the corresponding topological indicator.

2. Calculation of private desirables for each criterion in the interval of changes in the corresponding values of threshold coefficient from formula (11).

3. Calculation of the generalized Harrington desirability index as the geometric mean of all private desirables:

$$D = \sqrt[n]{\prod_{i=1}^{n} d_i}. \quad (15)$$

The maximum value of Harrington desirability index corresponds to the threshold coefficient, which makes it possible to obtain an optimal structure of a gene network based on the integrated analysis of topological parameters. Fig. 8 shows diagram of dependence of the comprehensive criterion on the threshold coefficient whose value changed in the interval from 0.45 to 0.55 with a step of 0.01.

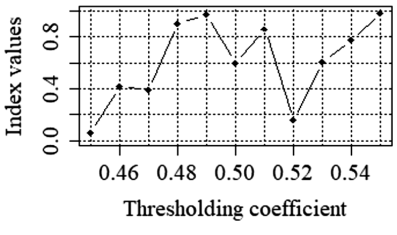

Fig. 8. Distribution diagram of the Harrington desirability index for different values of threshold coefficient

Analysis of the Figure shows that the optimal value based on individual parameters for the estimation of topology of a gene network is the value of threshold coefficient of 0.49. In this case, the network has 147 genes, centralization coefficient reaches a maximum, while clustering coefficient reaches a minimum. The values of coefficients of density and heterogeneity in the interval of change in the threshold coefficient from 0.45 to 0.49 monotonically changes towards lower and larger sides, respectively. In the interval from 0.49 to 0.51, the rate of change in these parameters is zero. The value of the comprehensive criterion, calculated based on the Harrington desirability function also reaches a maximum at a value of threshold coefficient of 0.49. Fig. 9 shows the result of reconstruction of a gene regulatory network when applying the algorithm of correlation output with a threshold coefficient of 0.49.

The research conducted allows us to propose a technology for the reconstruction of a gene regulatory network based on the correlation output algorithm. Structural block diagram of this technology is shown in Fig. 10.

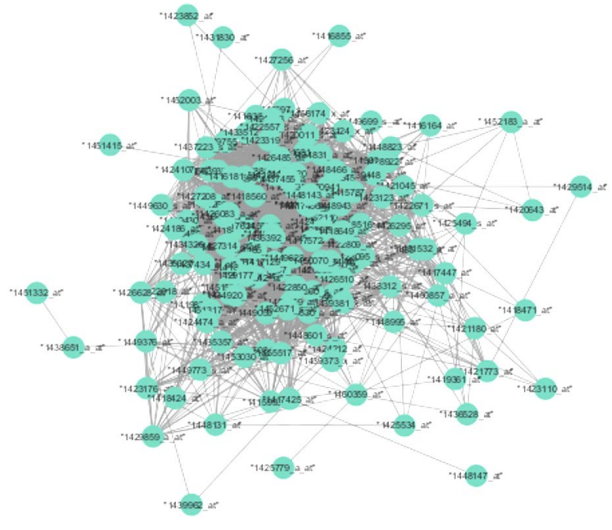

Fig. 9. Result of gene network reconstruction when applying the correlation output algorithm

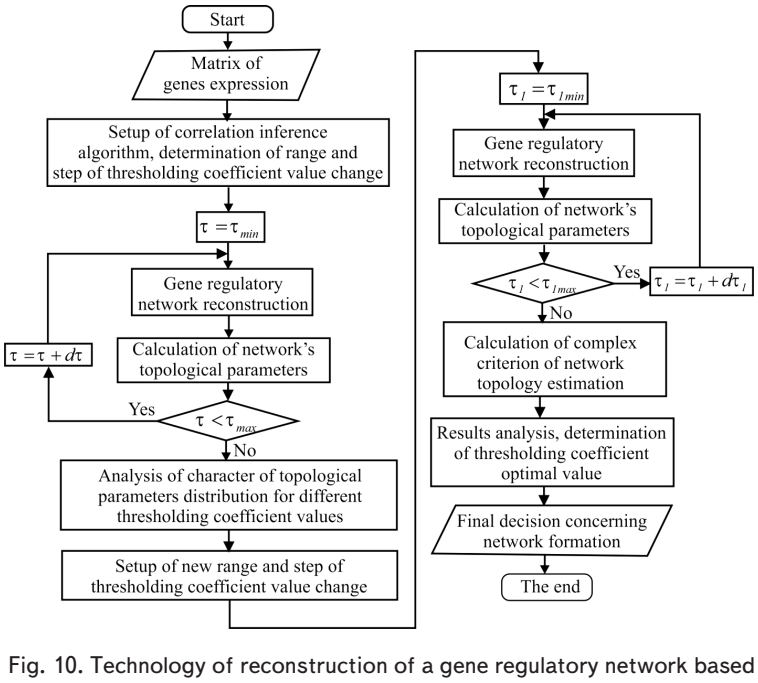

Fig. 10. Technology of reconstruction of a gene regulatory network based on the correlation output

Practical implementation of a given technology implies the following stages:

Stage I. Statement of problem. Formation of data.

1. Formation of initial data in the form of a matrix where lines are the genes; columns are the conditions for running a DNA-microchip experiment.

Stage II. Approximate estimation of the interval of change in threshold coefficient.

2. Assigning the approximate interval and step for a change in threshold coefficient. Initialization of the original value of threshold coefficient: $\tau=\tau_{min}$.

3. Reconstruction of GRN whose topology matches the assigned value of threshold coefficient.

4. Calculation of topological parameters of the obtained genetic regulatory network.

5. If a value of the threshold coefficient is less than the maximum value, we increase this value by $d\tau$ (step for a change in threshold coefficient) and proceed to step 3 of a given procedure. In the opposite case, we construct diagrams of dependence of the obtained topological parameters on the value of threshold coefficient.

6. Analysis of the results obtained; determining the new, narrower, interval and a smaller step for a change in the value of threshold coefficient.

Stage III. Determining the optimal value of threshold coefficient.

7. Reconstruction of a gene regulatory network within the new interval of change in the values of threshold coefficient. Calculation of parameters for the estimation of topology of a gene regulatory network at each step of change in the value of threshold coefficient.

8. Construction of diagrams for a change in the values of topological parameters depending on the threshold coefficient. Analysis of the obtained results. Determining the optimal value of threshold coefficient.

Stage IV. Reconstruction of a gene regulatory network.

9. Reconstruction of GRN, applying the optimal value of threshold coefficient.

6. Reconstruction of a gene regulatory network based on the algorithm ARACNE

The algorithm for the reconstruction of a gene regulatory network ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [24] forms a network topology based on the analysis of statistical hypotheses about the presence or absence of connection between relevant genes. As a result of the analysis, at the stage of network reconstruction, a vector of probabilities for each gene is derived, each element of which determines the presence and the strength of respective connection [25]:

$$P\left(\{g_i\}\right)= \\ =\frac{1}{Z}\exp\left[-\sum_{i=1}^{N}\phi_i\left(g_i\right)-\sum_{i,j=1}^{N}\phi_{i,j}\left(g_i,g_j\right)-\sum_{i,j,k=1}^{N}\phi_{i,j,k}\left(g_i,g_j,g_k\right)-...\right], \quad (16)$$

where N is the number of genes, Z represents a normalizing factor, ϕ are the potentials that determine the strength of connection of the respective group of genes.

It is believed that the sets of genes interact with each other if the corresponding potential is different from zero. Otherwise, there is no connection between a given group of genes. Correct choice of parameters for a given algorithm makes it possible to obtain a gene network with significantly fewer connections compared with the network obtained when applying the correlation output algorithm, which simplifies the interpretation of results during subsequent modeling of GRN. Assessment of the degree of relationship between the pair of genes g_i, g_j in the presence of M options for connection is obtained using the Gaussian nuclear assessment based on Shannon entropy [25]:

$$I\left(\{g_i\},\{g_j\}\right)=\frac{1}{M}\sum_{k=1}^{M}\log\frac{H_k\left(g_i,g_j\right)}{H_k\left(g_i\right)H_k\left(g_j\right)}, \quad (17)$$

where $H(g)$ is the Shannon entropy, which is calculated for the profile of gene g. The basic idea of ARACNE algorithm lies in the fact that in the presence of various ways of connection in the network, each of which is characterized by the degree of appropriate relationship $I(g_i, g_j)$, the connection is chosen that satisfies condition:

$$I\left(g_i,g_j\right)<\min\left[I\left(g_i,g_s\right),I\left(g_s,g_p\right),...,I\left(g_h,g_j\right)\right], \quad (18)$$

where g_s, g_p,..., g_h are the intermediate genes that carry the relationship between genes g_i and g_j. This optimizes the number of connections in the network. As a result, we obtain a network of interacting genes, the weight of the connection between the relevant genes in which is determined by the degree of connection between the genes. The number of network connections is limited also by the introduction of a threshold coefficient. It is believed that if the weight of the corresponding connection is less than the value of the threshold coefficient, the link between these genes is broken.

Simulation of the process of a gene network reconstruction based on the output algorithm ARACNE was also performed in the programming environment *CytoScape* using data on the profiles of gene expressions *moe430a* from the database *ArrayExpress*. According to the technology for determining optimal value of a threshold coefficient (Fig. 10), at the first stage the value of threshold coefficient varied in the interval from 0.1 to 0.9 with a step of 0.1. Distribution diagrams of individual topological network parameters depending on the threshold coefficient are shown in Fig. 11. A network clustering coefficient was equal to zero, indicating the absence of links between the neighbors of genes in the network. Based on the analysis of the obtained results we can conclude that the optimal value of the threshold coefficient is in the range from 0.3 to 0.5 because the values of coefficients of centralization and heterogeneity in a given range reach local maxima and the density of nodes – a local minimum. The number of genes thus varies from 146 to 147, which is quite acceptable.

Fig. 12 shows similar diagrams for the case of change in the value of threshold coefficient from 0.3 to 0.5 with a step of 0.02. An analysis of the acquired diagrams does not make it possible to uniquely select the optimal value of threshold coefficient, because coefficients of centralization, heterogeneity and density have three local extrema, which to some extent contradict each other. In line with the technique for determining an optimal value of threshold coefficient, at the final step we calculated the integrated criterion based on the Harrington desirability function, which contained as a component the corresponding topological parameters. A diagram of dependence of the value of comprehensive criterion on the threshold coefficient is shown in Fig. 13. An analysis of Fig. 13 reveals that the maximum value of the generalized Harrington desirability index is reached at threshold coefficients of 0.33 and 0.42. However, it should be noted that in the second case the genetic network is less by five genes, which is why the value of threshold coefficient at 0.33 is more acceptable in

terms of the number of genes in the network. Fig. 14 shows the result of a gene network reconstruction when applying the value of threshold coefficient of 0.33.

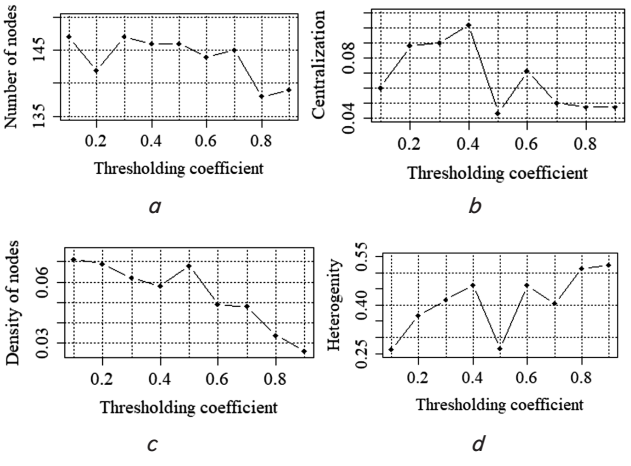

Fig. 11. Distribution diagrams of topological parameters when changing the threshold coefficient from 0.1 to 0.9 with a step of 0.1 when using the ARACNE algorithm: *a* – number of genes; *b* – centralization coefficient; *c* – density; *d* – heterogeneity

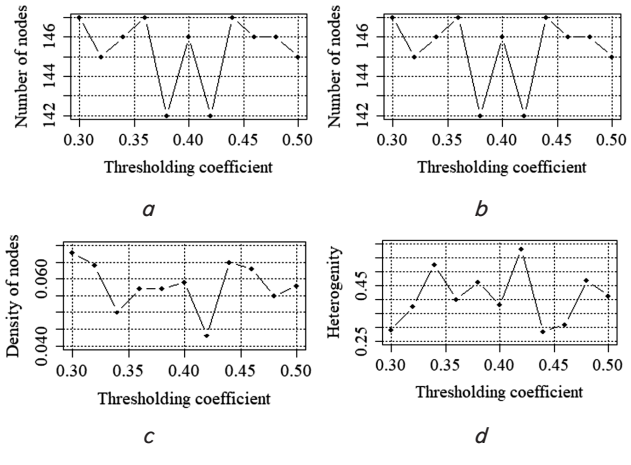

Fig. 12. Distribution diagrams of topological parameters when changing the threshold coefficient from 0.3 to 0.5 with a step of 0.02 when using the ARACNE algorithm: *a* – number of genes; *b* – centralization coefficient; *c* – density; *d* – heterogeneity

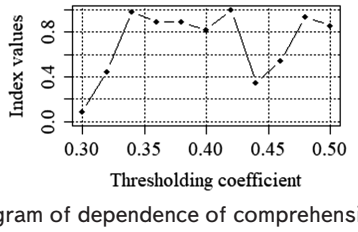

Fig. 13. Diagram of dependence of comprehensive criterion for the estimation of a network topology on the threshold coefficient

Practical implementation of the proposed technology for the reconstruction of a gene network allows us to optimize the topology of the network, but there is a problem on the validation of GRN obtained. This problem can be solved using a ROC analysis (Receiver Operator Characteristic) [26] that is applied to visualize results of binary classification using errors of the first and second kind.

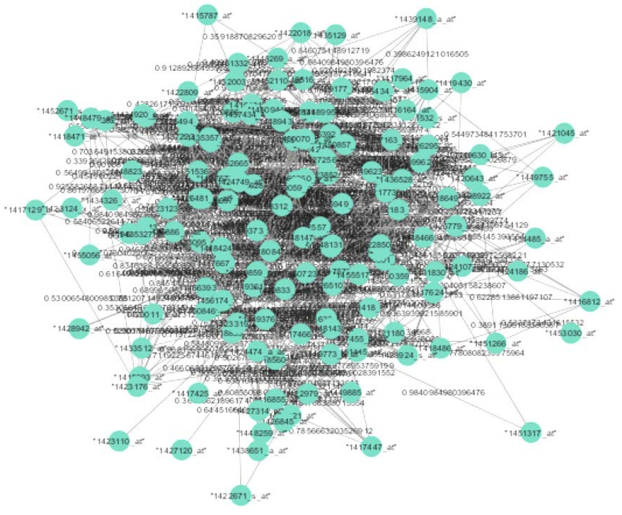

Fig. 14. Result of the reconstruction of a gene network when applying the algorithm ARACNE

7. Discussion of results: validation technique of the model of a genetic regulatory network

The process of validation of GRN model implies a comparative analysis of the character of relationships between genes in the network reconstructed based on a selected group of genes and conditions compared with the network reconstructed on the basis of all examined genes and conditions. Gene networks are considered to be completely adequate if the character of relations between relevant genes in different networks fully coincides. In this case, we estimate the presence of a relationship between genes. If there is a connection, it is considered to be equal to 1; in the absence of such a connection, this value is 0. According to the theory of ROC analysis, at the first stage we calculate quality parameters for the classification of relationships between genes in relevant networks. Such indicators are:

– TP (True Positives) – the number of relationships between pairs of matching genes that coincide in the two networks (true positive cases).

– TN (True Negatives) – the number of matching negative relationships between pairs of corresponding genes in different networks (true negative cases).

– FN (False Negatives) – the number of relationships between pairs of genes, reconstructed based on full data not identified in the network, reconstructed on the basis of a limited number of genes and conditions (error of the first kind). In this case, the relationship that exists between a pair of genes in the full network is missing between the pair of genes in the examined network.

– FP (False Positives) – the number of missing links between the relevant genes in the network, reconstructed based on full data that are identified as existing in the network, reconstructed on the basis of a limited number of genes and conditions (error of the second kind). In this case, a connection between a pair of genes that is missing in the full network is identified as existing between a given pair of genes in the examined network.

Based on these parameters, we calculate relative indicators for the model quality assessment:

– the percentage of true positive cases or the sensitivity of the model – the ratio of the number of true positive connections to the full number of connections between the examined genes, based on the results of analysis of the gene network employing complete data:

$$Sc = TPR = \frac{TP}{TP + FN} \cdot 100\,\%. \quad (19)$$

– Percentage of false positive cases:

$$FPR = \frac{FP}{TN + FP} \cdot 100\,\%. \quad (20)$$

– Specificity – the percentage of missing links that were correctly identified by the network, reconstructed based on a limited number of genes and conditions:

$$Sp = \frac{TN}{TN + FP} \cdot 100\,\%. \quad (21)$$

It should be noted that the percentage of false positive cases and the specificity are related via ratio: $FPR=100-Sp$. A larger specificity value corresponds to a smaller percentage of incorrectly identified cases of the presence of links in the complete network. *ROC*-curve is a dependence diagram of sensitivity *Sc* on the percentage of incorrect positive cases $FPR=100-Sp$. A larger value of sensitivity and a lower *FPR* value corresponds to a higher degree of the adequacy of a model. In this case, the area under a *ROC*-curve (AUC) reaches the highest value. Another criterion that determines the adequacy of a model is calculated as the ratio of sensitivity to the percentage of false positive cases:

$$RC = \frac{Sc}{FPR}. \quad (22)$$

A higher value of this criterion corresponds to a greater level of adequacy of the gene network, reconstructed based on the corresponding bicluster, the gene network based on the totality of the genes and conditions. A structural block diagram of validation technique of a gene network is shown in Fig. 15.

Practical implementation of a given technique implies the following steps:

1. Statement of problem. Forming an array of gene expression profiles. Data preprocessing: filtering, reduction, clustering of gene expression profiles.

2. Reconstruction of basic gene networks based on data from obtained clusters.

3. Biclustering of data on gene expressions contained in the derived clusters. Fixing the biclusters.

4. Reconstruction of gene networks based on data about gene expressions of the relevant biclusters.

5. Determining the quality indicators for the classification of relationships between genes in the derived networks (*TP, TN, FP, FN*).

6. Calculation of relative indicators for the model quality estimation *Sc, Sp, FPR* according to formulae (19)–(21).

7. Construction of dependence diagram of sensitivity *Sc* on the percentage of false positive cases *FPR*.

8. Selection of the model whose area under the ROC curve is the largest or which corresponds to the higher value of the relative criterion calculated from formula (22).

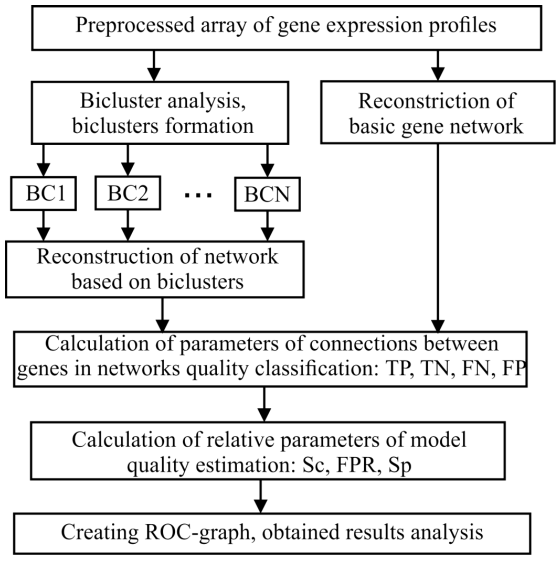

Fig. 15. Structural block diagram of validation technique for a gene network

7. 1. Validation of the model of a gene network based on the correlation inference algorithm

Fig. 16 shows results of the biclustering analysis of gene expression profiles data *moe430a*, performed using the algorithm "ensemble" [26].

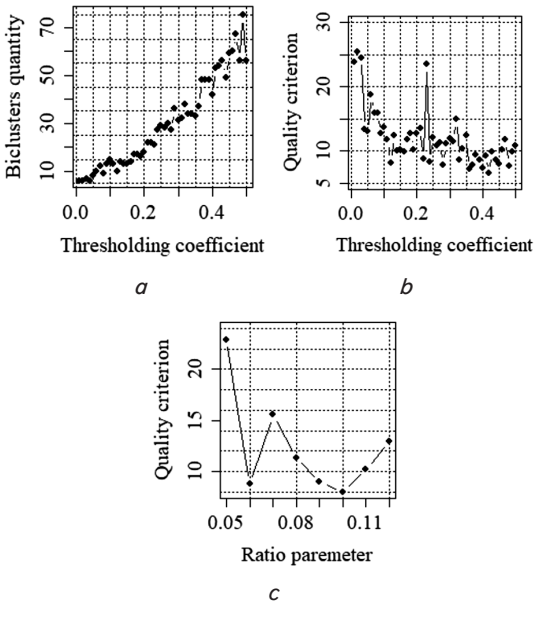

Fig. 16. Results of the biclustering analysis of gene expression profiles data *moe430a*: *a* – dependence diagram of the number of biclusters on the threshold coefficient; *b* – dependence diagram of the criterion of biclustering quality on the threshold coefficient; *b* – dependence diagram of the criterion of biclustering quality on relative threshold coefficient

At the first stage, the value of a parameter that determines the ratio of the number of lines and columns in biclusters was fixed at the level of 0.15, the value of threshold

coefficient changed from 0.01 to 0.5 with a step of 0.01. The value of the threshold coefficient based on the analysis of the diagram shown in Fig. 16, *b* was fixed at the level of 0.12 (the first global minimum). Increasing the value of a given criterion is not appropriate because this would lead to an increase in the number of small biclusters. The value of relative threshold coefficient changed in the range from 0.05 to 0.12 with a step of 0.01. Fig. 16 shows that the minimum value of the internal criterion of biclustering quality corresponds to the value of relative threshold coefficient of 0.1.

The result of a biclustering analysis of gene expression profiles using the biclustering algorithm "ensemble", at a threshold coefficient of 0.12 and a relative threshold coefficient of 0.1, is given in Table 2. Lines denote the number of genes in a bicluster, columns – the number of conditions for determining the gene expression profiles.

Table 2

Distribution of lines and columns in the biclusters derived for the gene expression profiles data *moe430a*

BC	1	2	3	4	5	6	7	8	9	10	11	12	13
Lines	23	16	9	13	5	39	11	44	32	28	24	24	6
Columns	8	8	4	9	8	8	7	12	12	6	11	9	6

To validate the models of gene regulatory networks reconstruction, we selected biclusters, which contain more than ten genes (small biclusters were not dealt with). Thus, 10 biclusters were chosen: BC1, BC2, BC4, BC6–BC12. Dependence diagrams of values of the generalized Harrington desirability index on threshold coefficient for the models of gene networks based on the obtained biclusters are shown in Fig. 17. The threshold coefficient changed from 0.35 to 0.55 with a step of 0.01. This interval was determined empirically. The value of threshold coefficient in the specified interval corresponded to the complete number of genes in the obtained networks.

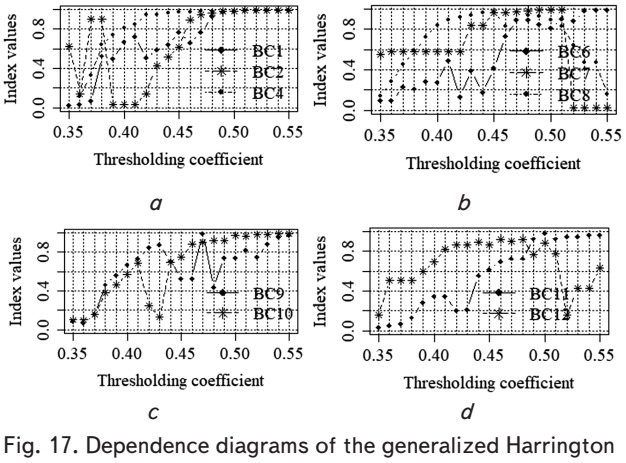

Fig. 17. Dependence diagrams of the generalized Harrington desirability index on value of the threshold coefficient for gene networks, reconstructed based on data from biclusters using the correlation output algorithm:
a — biclusters 1, 2, 4; *b* — biclusters 6, 7, 8;
c — biclusters 9, 10; *d* — biclusters 11, 12

Based on an analysis of the obtained diagrams, we registered the following threshold coefficients: BC1, BC2 and BC7 – 0.51; BC4 and BC9 – 0.47; BC6 – 0.53; BC8 – 0.45; BC10 and BC11 – 0.5; BC12 – 0.48. Fig. 18 shows ROC curves for the derived models and values of the relative estimation criteria for the adequacy of models of gene networks, reconstructed using data from the respective biclusters. Horizontal line in Fig. 18, *b* is drawn at the level of the average relative criterion of estimation of the adequacy of the model for the derived gene networks. An analysis of the results obtained showed that the value for the specificity parameter is in the range from 97 to 100 percent, indicating a low percentage of incorrectly identified positive cases. The value of sensitivity for the derived biclusters varies from 45.5 % for the gene network based on data of the seventh bicluster, to 90.2 % for the network, reconstructed on the basis of data from the fourth bicluster. The minimum value of the relative criteria, calculated from formula (22), corresponds to the gene network based on the eighth bicluster and equals 21. In this case, the maximum value of this criterion at 1,746 matches the fourth bicluster. The weighted average of relative criterion for the validation of derived models equals 355.5.

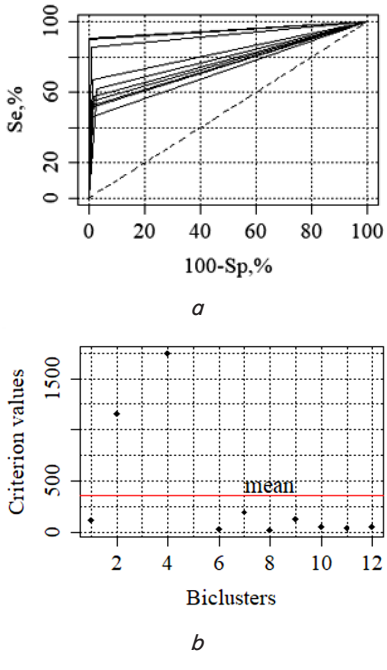

Fig. 18. Validation results of technique for the reconstruction of gene networks based on the correlation output algorithm:
a — ROC curves for the obtained GRN models;
b — distribution diagram of the values of relative criterion for the validation of obtained biclusters

The results obtained indicate a high level of adequacy of the proposed technique for the reconstruction of gene networks as the values of relative validation criteria for all reconstructed gene networks are substantially larger than unity. The number of incorrectly identified positive cases belongs to the interval from 0 to 3 percent, and sensitivity is less than 50 percent (45.5) only for the network based on the seventh bicluster. For the gene networks based on other biclusters, the value of a given parameter is greater than 50 percent, and for the fourth bicluster it reaches 90.2 percent.

7. 2. Validation of the model of a genetic network based on the algorithm ARACNE

Fig. 19 shows dependence diagrams of values of the generalized Harrington desirability index on threshold coeffi-

cient for the models of gene networks based on the obtained biclusters applying the algorithm ARACNE. Threshold coefficient changed from 0.03 to 0.2 with a step of 0.01. This interval was also determined empirically. The value of threshold coefficient in the specified interval corresponded to the complete number of genes in the obtained networks. Based on an analysis of the obtained diagrams, we registered the following values of threshold coefficients: BC1, BC4, BC9 and BC12 – 0.13; BC2 – 0.06; BC6 – 0.19; BC7 – 0.07; BC8 – 0.17; BC10 – 0.14 and BC11 – 0.09. Fig. 20 shows ROC curves for the derived models and values for relative estimation criteria of adequacy of models of the gene networks, reconstructed using data from the respective biclusters.

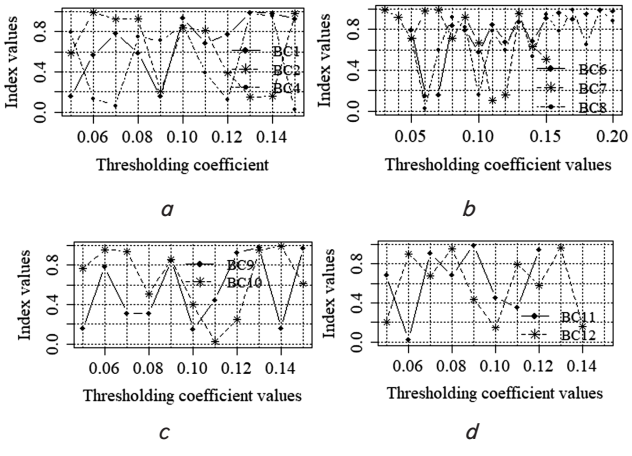

Fig. 19. Dependence diagrams of the generalized Harrington desirability index on the value of threshold coefficient for the gene networks, reconstructed on the basis of data from biclusters using the algorithm ARACNE: *a* – biclusters 1, 2, 4; *b* – biclusters 6, 7, 8; *c* – biclusters 9, 10; *d* – biclusters 11, 12

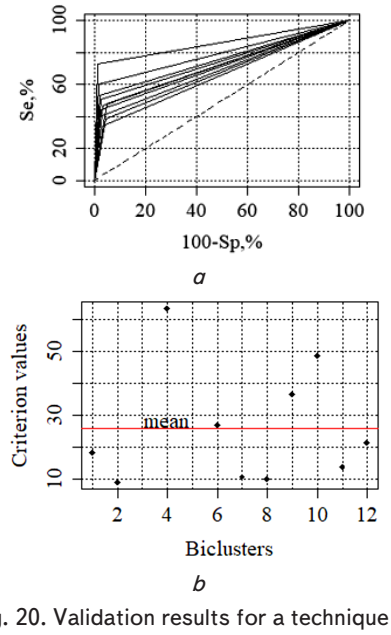

Fig. 20. Validation results for a technique for the reconstruction of gene networks based on the algorithm ARACNE: *a* – ROC curves for the obtained GRN models; *b* – distribution diagram of the values of relative validation criterion for the obtained biclusters

An analysis of the obtained results allows us to conclude that the level of adequacy of the models of gene networks, reconstructed using the output algorithm ARACNE is significantly lower compared with the networks, reconstructed using the output correlation algorithm. Thus, in the case of application of the algorithm ARACNE, sensitivity varies from 34.3 % to 72.5 %, and specificity – ranging from 95.1 % to 99.2 %. The weighted average of the relative validation criteria for the models of gene networks based on the output algorithm ARACNE is equal to 25.24, significantly less than the corresponding value when applying the algorithm of correlation output. This indicates a higher output correlation algorithm efficiency in comparison with the algorithm for reconstruction of gene networks ARACNE.

8. Conclusions

1. We have developed a technique for the reconstruction of gene regulatory networks based on comprehensive application of network topological parameters and the generalized Harrington desirability index. The technique is given in the form of a structural block diagram for a stepwise process of information processing for determining optimal parameters of the algorithm, applied for the reconstruction of a gene network. We obtained dependence diagrams of topological parameters and the generalized Harrington desirability index on the value of a threshold coefficient. It is shown that when the correlation output algorithm is employed, the optimal network topology is achieved at a value of threshold coefficient of 0.49. In this case, the network has 147 genes, the centralization coefficient reaches a maximum while the clustering coefficient – a minimum. Values of the coefficients of density and heterogeneity in the range of change in threshold coefficient of 0.45 to 0.49 monotonically change toward lower and larger sides. Value of the comprehensive criterion, calculated on the basis of the Harrington desirability function, also reaches a maximum at a value of threshold coefficient of 0.49.

2. We have developed a technique for the validation of models of gene networks based on ROC analysis, the implementation of which implies a comparative analysis of the character of relations between relevant genes in the network based on the totality of genes and gene networks on the basis of the obtained biclusters. The process of validation implies determining errors of the first and second kind, with subsequent calculation of the relative criterion, derived as the ratio of sensitivity of the model to the percentage of false positive cases. A larger value of this criterion corresponds to a higher level of adequacy of the respective model. It is shown that when applying the correlation output algorithm, the value of specificity parameter is in the range from 97 to 100 percent, indicating a low percentage of incorrectly identified positive cases. The value of sensitivity for the derived biclusters varies from 45.5 % for the gene network based on data from the seventh bicluster, to 90.2 % for the network, reconstructed on the basis of data from the fourth bicluster. Minimum value of a relative criterion for the model validation quality corresponds to the gene network based on the eighth bicluster and equals 21. In this case, the maximum value of this criterion of 1,746 matches the fourth bicluster. The weighted average of relative validation criterion for the derived models

equals 355.5. When employing the algorithm ARACNE, the sensitivity varies from 34.3 % to 72.5 %, and the specificity – from 95.1 % to 99.2 %. The weighted average of relative validation criteria for the models of gene networks based on the output algorithm ARACNE equals 25.24.

3. Simulation was performed of the processes of reconstruction and validation of the gene networks, obtained using the algorithms of correlation output and ARACNE. We employed, as model data, the gene expression profiles data *moe430a* from the database *ArrayExpress*. The data were acquired in the course of DNA-microchip experiments and contained information on gene expression of mesenchymal cells of two types: nerve crest and mesoderm. The matrix of output data consisted of 147 lines, or genes, and 20 columns, or conditions for determining expressions of the relevant genes. We have obtained optimal topologies of the corresponding gene networks that match the maximum value of the generalized Harrington desirability index. It is shown that in terms of analysis of the character of relations between relevant genes, the algorithm of output correlation is more effective compared to the algorithm ARACNE. The weighted average of a relative validation criterion for the derived models using the correlation output algorithm is equal to 355.5, which considerably exceeds the respective value of 25.24 when applying the algorithm ARACNE.

References

1. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse / Liu Z.-P., Wu C., Miao H., Wu H. // Database. 2015. Vol. 2015. P. bav095. doi: 10.1093/database/bav095
2. Protein complex analysis: From raw protein lists to protein interaction networks / Meysman P., Titeca K., Eyckerman S., Tavernier J., Goethals B., Martens L. et. al. // Mass Spectrometry Reviews. 2015. Vol. 36, Issue 5. P. 600–614. doi: 10.1002/mas.21485
3. Linear modeling of mRNA expression levels during CNS development and injury / D'haeseleer P., Wen X., Fuhrman S., Somogyi R. // Pacific Symposium on Biocomputing. 1999. P. 41–52.
4. Liang S., Fuhrman S., Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures // Pacific Symposium on Biocomputing. 1998. P. 18–29.
5. Using Bayesian Networks to Analyze Expression Data / Friedman N., Linial M., Nachman I., Pe'er D. // Journal of Computational Biology. 2000. Vol. 7, Issue 3-4. P. 601–620. doi: 10.1089/106652700750050961
6. Chen T., He H. L., Church G. M. Modeling gene expression with differential equations // Proceedings of the Pacific Symposium on Biocomputing. 1999. P. 29–40.
7. Wong K.-C., Li Y., Zhang Z. Unsupervised Learning in Genome Informatics // Unsupervised Learning Algorithms. 2016. P. 405–448. doi: 10.1007/978-3-319-24211-8_15
8. Boolean modeling techniques for protein co-expression networks in systems medicine / Mayer G., Marcus K., Eisenacher M., Kohl M. // Expert Review of Proteomics. 2016. Vol. 13, Issue 6. P. 555–569. doi: 10.1080/14789450.2016.1181546
9. Emmert-Streib F., Dehmer M., Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks // Frontiers in Cell and Developmental Biology. 2014. Vol. 2. doi: 10.3389/fcell.2014.00038
10. Wang K., Zhang L., Liu X. A review of gene and isoform expression analysis across multiple experimental platforms // Chinese Journal of Biomedical Engineering. 2017. Vol. 36, Issue 2. P. 211–218.
11. An integrative method to decode regulatory logics in gene transcription / Yan B., Guan D., Wang C., Wang J., He B., Qin J. et. al. // Nature Communications. 2017. Vol. 8, Issue 1. doi: 10.1038/s41467-017-01193-0
12. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease / Alkallas R., Fish L., Goodarzi H., Najafabadi H. S. // Nature Communications. 2017. Vol. 8, Issue 1. doi: 10.1038/s41467-017-00867-z
13. Nair A., Chetty M., Wangikar P. P. Improving gene regulatory network inference using network topology information // Molecular BioSystems. 2015. Vol. 11, Issue 9. P. 2449–2463. doi: 10.1039/c5mb00122f
14. Detection of Complexes in Biological Networks Through Diversified Dense Subgraph Mining / Ma X., Zhou G., Shang J., Wang J., Peng J., Han J. // Journal of Computational Biology. 2017. Vol. 24, Issue 9. P. 923–941. doi: 10.1089/cmb.2017.0037
15. Identification of Protein Complexes by Integrating Multiple Alignment of Protein Interaction Networks / Ma C.-Y., Phoebe Chen Y.-P., Berger B., Liao C.-S. // Bioinformatics. 2017. P. btx043. doi: 10.1093/bioinformatics/btx043
16. Pontes B., Giráldez R., Aguilar-Ruiz J. S. Biclustering on expression data: A review // Journal of Biomedical Informatics. 2015. Vol. 57. P. 163–180. doi: 10.1016/j.jbi.2015.06.028
17. Criterial analysis of gene expression sequences to create the objective clustering inductive technology / Babichev S., Taif M. A., Lytvynenko V., Osypenko V. // 2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO). 2017. doi: 10.1109/elnano.2017.7939756
18. Gene expression sequences clustering based on the internal and external clustering quality criteria / Babichev S., Krejci J., Bicanek J., Lytvynenko V. // 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2017. doi: 10.1109/stc-csit.2017.8098744
19. Objective Clustering Inductive Technology of Gene Expression Sequences Features / Babichev S., Lytvynenko V., Korobchynskyi M., Taiff M. A. // Communications in Computer and Information Science. 2017. P. 359–372. doi: 10.1007/978-3-319-58274-0_29
20. Model of the Objective Clustering Inductive Technology of Gene Expression Profiles Based on SOTA and DBSCAN Clustering Algorithms / Babichev S., Lytvynenko V., Skvor J., Fiser J. // Advances in Intelligent Systems and Computing. 2017. P. 21–39. doi: 10.1007/978-3-319-70581-1_2
21. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks // Genome Research. 2003. Vol. 13, Issue 11. P. 2498–2504. doi: 10.1101/gr.1239303
22. Computing topological parameters of biological networks / Assenov Y., Ramírez F., Schelhorn S.-E., Lengauer T., Albrecht M. // Bioinformatics. 2007. Vol. 24, Issue 2. P. 282–284. doi: 10.1093/bioinformatics/btm554

23. Neural crest and mesoderm lineage-dependent gene expression in orofacial development / Bhattacherjee V., Mukhopadhyay P., Singh S., Johnson C., Philipose J. T., Warner C. P. et. al. // Differentiation. 2007. Vol. 75, Issue 5. P. 463–477. doi: 10.1111/j.1432-0436.2006.00145.x
24. Harrington J. The desirability function // Industrial Quality Control. 1965. Vol. 21, Issue 10. P. 494–498.
25. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context / Margolin A. A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera R., Califano A. // BMC Bioinformatics. 2006. Vol. 7, Issue (Suppl 1). P. S7. doi: 10.1186/1471-2105-7-s1-s7
26. Fawcett T. ROC graphs: Notes and practical considerations for researches // Machine learning. 2004. Vol. 31, Issue 1. P. 1–38.
27. Kaiser S. Biclustering: Methods, software and application. Munchen, 2011. 178 p.

Пропонуються адаптивні комбіновані моделі гібридного та селективного типів для прогнозування часових рядів на основі програмного набору з адаптивних поліноміальних моделей різних порядків. Пропонуються адаптивні комбіновані моделі прогнозування часових рядів з врахуванням результатів ідентифікації подібностей в ретроспекції цих часових рядів. Оцінена ефективність прогнозування різних комбінованих моделей залежно від рівня персистентності часових рядів. Розроблені моделі дозволяють підвищити точність у випадку середньострокового прогнозування нестаціонарних часових рядів, зокрема фінансових показників

Ключові слова: прогнозування часових рядів, пошук подібностей, адаптивна комбінована модель, показник Герста

Предлагаются адаптивные комбинированные модели гибридного и селективного типов для прогнозирования временных рядов на основе программного набора из адаптивных полиномиальных моделей разных порядков. Предлагаются адаптивные комбинированные модели прогнозирования временных рядов с учетом результатов идентификации подобий в ретроспекции этих временных рядов. Оценена эффективность прогнозирования различных комбинированных моделей в зависимости от уровня персистентности временных рядов. Разработанные модели позволяют повысить точность в случае среднесрочного прогнозирования нестационарных временных рядов, в частности финансовых показателей

Ключевые слова: прогнозирование временных рядов, поиск подобий, адаптивная комбинированная модель, показатель Херста

UDC 004:519.2
DOI: 10.15587/1729-4061.2018.121620

DEVELOPMENT OF ADAPTIVE COMBINED MODELS FOR PREDICTING TIME SERIES BASED ON SIMILARITY IDENTIFICATION

A. Kuchansky
PhD, Associate Professor
Department of Cybersecurity and Computer Engineering*
E-mail: kuczanski@gmail.com
A. Biloshchytskyi
Doctor of Technical Sciences, Professor, Head of Department
Department of Information Systems and Technologies
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033
Yu. Andrashko
Lecturer
Department of System Analysis and Optimization Theory
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000
S. Biloshchytska
PhD, Associate Professor
Department of Information Technology Designing and Applied Mathematics*
Ye. Shabala
PhD, Associate Professor
Department of Cybersecurity and Computer Engineering*
O. Myronov
Postgraduate student
Department of Information Technologies*
*Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03037

1. Introduction

It is known that the overwhelming majority of financial, technical and physical processes for which the problem of predicting arises are characterized by nonlinearity and instability with respect to the average level. Application of classical econometric prediction models and appropriate methods of predicting such time series that reflect these processes is rather limited. This is because of a low efficiency of these models in such conditions. Prospective directions