IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction

S. Babichev
Kherson National Technical University
Kherson, Ukraine
Jan Evangelista Purkyne University
Usti nad Labem, Czech Republic
sergii.babichev@ujep.cz

V. Lytvynenko
Kherson National Technical University
Kherson, Ukraine
immun56@gmail.com

J. Škvor2
Jan Evangelista Purkyne University
Usti nad Labem, Czech Republic
jskvor@physics.ujep.cz

M. Korobchynskyi
Military-Diplomatic Academy named Eugene Bereznyak
Kiev, Ukraine
maks_kor@ukr.net

M. Voronenko
Kherson National Technical University
Kherson, Ukraine

***Abstract*—The paper presents the information technology of gene expression profiles processing in order to reconstruct gene regulatory networks. The information technology is presented as a structural block-chart, which contains all stages of studied data processing. DNA microchips of patients, which were studied on different types of diseases, were used as experimental data. The relative criteria of validation for all reconstructed networks were calculated during simulation process. The obtained results show high efficiency of the proposed technology. High values of the validation criteria indicate a high level of the obtained gene networks objectivity.**

Keywords—objective clustering, reduction, biclustering gene expression profiles, gene regulatory network, reconstruction, validation

I. INTRODUCTION

Actuality of the problem is determined by the modern state of works in the field of gene expression profiles processing for the purpose of gene regulatory networks reconstruction. Gene regulatory network is a set of genes, which interact with each other to control the specific cell functions [1]. Qualitatively reconstructed gene regulatory network promotes to better understanding of the gene interaction mechanism in order to create new methods to early diagnostics and treatment of complex genetic diseases. The gene expression profiles, which are obtained by DNA microarray experiments or by RNA sequences technology, are the basis for the reconstruction of gene regulatory networks [2, 3]. High dimension of feature space is one of the distinctive peculiarities of the studied data. The reconstruction of gene networks based on the whole dataset of gene expression profiles is very complicated task/due to the following aspects: it requests large computer resources; complexity of the obtained networks complicates the obtained results interpretation. Therefore, it is necessary at early stage of gene regulatory network reconstruction to process the gene expression profiles with the use of current computational and information technologies of complex data processing. This process includes data filtering in the case of DNA microchip experiment performing, non-informative genes reducing, data clustering and biclustering in order to select mutually correlated genes and samples.

The issues concerning creation of the hybrid clustering methods were investigated in [4,5]. The authors propose neural network, which allows them to increase the quality of information processing. In [6,7] the authors proposed system which solves a clustering task of non-stationary data streams under uncertainty conditions when data come in the form of a sequential stream in an online mode. However, it should be noted that authors' researches are primarily focused on low-dimensional data processing. High-dimensional data processing are not considered in these works. In the papers [8, 9] the author presented a novel approach to solving the problem of task allocation among the agents which takes into account the restrictions on agents' communications and self-diagnosis strategies for multinodular systems. In [10–12] the authors considered the issues of handling uncertainties in the problems of modeling and forecasting dynamic systems within the framework of the dynamic planning methodology. However, the proposed methods do not allow us to increase the efficiency of the genes and conditions grouping. Bicluster analysis is actual to solve this problem nowadays [13,14]. The rows and columns are grouping in accordance with their mutual correlation during biclustering process. One of the main disadvantages of this technology is a high percent of information lost due to the high dimension of the initial data array. However, it should be noted that the effective technology of gene expression profiles processing does not exist nowadays. This fact can be explained by high dimension of features space that requests the use of complex data processing modern techniques.

The Aim of the paper is the development of information technology of gene expression profiles processing in order to reconstruct gene regulatory networks.

II. INFORMATION TECHNOLOGY OF GENE EXPRESSION PROFILES PROCESSING

The structural block-chart of the information technology of gene expression profiles processing in order to reconstruct gene expression networks is presented in Fig. 1.

The implementation of this technology involves the following stages:

Stage I. Formation of gene expression profiles array in the case of DNA microchip experiments.

The matrix of light intensities is obtained during DNA microchip experiment performing. Firstly, it is necessary to transform the light intensities to the expression of the corresponding genes. Implementation of this stage involves the following steps: background correction, normalization, PM correction and summarization. The methods, which can be used at each step, are shown in Fig. 2.

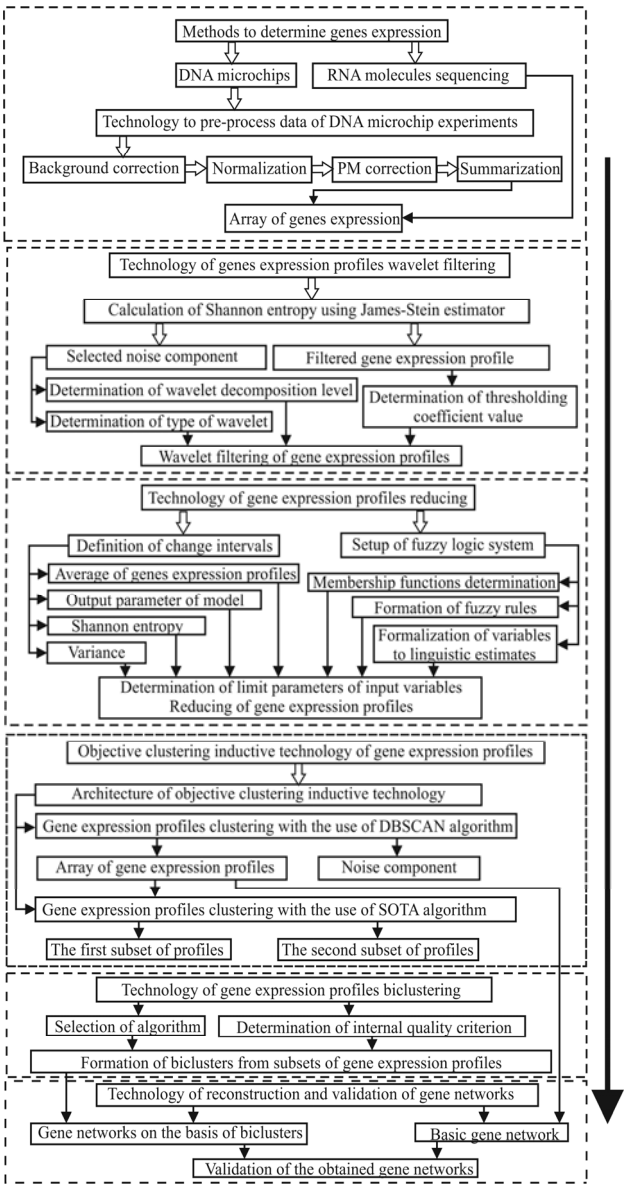

Fig. 1. Information technology of gene expression profiles processing

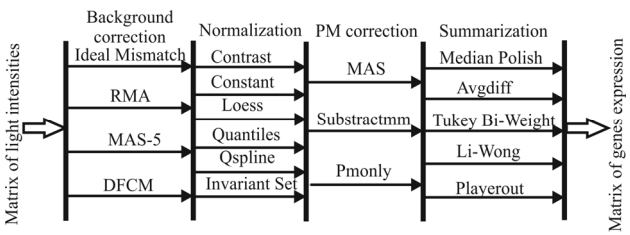

Fig. 2. Methods to evaluate the expression of genes

Estimation of data processing quality was performed with the use of Shannon entropy criterion, which was calculated based on James-Stein shrinkage estimator [15]. This method is based on the complex use of two different models: a high-dimensional model with low bias and high variance, and a lower dimensional model with larger bias but smaller variance. The probability of values distribution in a cell is calculated as follows:

$$p_i^{Shrink} = \lambda p_i + (1-\lambda)\, p_i^{ML} \qquad (1)$$

where p_i^{ML} is the probability of data values distribution in *i-th* cell, which is calculated by the maximum likelihood method, $p_i = 1/n_i$ is the estimation of probability in *i*-th cell, n_i is the quantity of features in this cell. Obviously, that p_i^{ML} corresponds to the high-dimensional model with low bias and high variance and p_i is the estimation with higher bias and lower variance of the features distribution. Intensity parameter λ in the proposed model is calculated as follows:

$$\lambda = \frac{1-\sum_{i=1}^{k}\left(p_i^{ML}\right)^2}{(n-1)\sum_{i=1}^{k}\left(p_i - p_i^{ML}\right)^2}, \qquad (2)$$

where n is the features quantity in the studied vector. Shannon entropy value in this case is estimated with the use of standard formula taking into account the method of probability calculation in the appropriate cell:

$$H^{Shrink} = -\sum_{i=1}^{k} p_i^{Shrink} \log_2 p_i^{Shrink} \qquad (3)$$

It is obvious, that in the case of gene expression profiles informativity evaluation, the minimum value of Shannon entropy criterion corresponds to higher quality of the investigated data processing. Choice of the optimal combination of the methods was performed based on the minimum value of Shannon entropy during the enumeration of all combinations of these methods.

Stage II. Wavelet filtering of gene expression profiles.

The necessity of this stage is determined by the existence of background noise, which can be appeared during scanning of information from DNA microchip. The proposed technology of wavelet filter optimal parameters determination involves concurrent evaluation of Shannon entropy for both the filtered data and allocated noise component. Structural flowchart of this process is presented in Fig. 3.

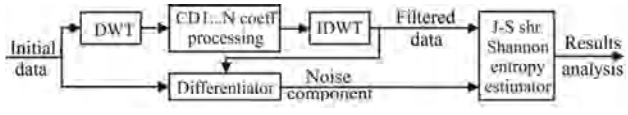

Fig. 3. Structural flowchart of wavelet-filtering process

Implementation of this process involves the following steps:

1. Choice of the mother wavelet from the list of the available wavelets.

2. Determination of the wavelet decomposition optimal level for the studied profiles based on the maximum

value of Shannon entropy, which is calculated for the allocated noise component. At this step the choice of the wavelet type from the family of the mother wavelet and the value of the thresholding coefficient are setup randomly from the range of the available values.

3. Determination of the wavelet type from the family of the appropriate mother wavelet based on the maximum value of Shannon entropy, which is calculated for the allocated noise component.

4. Determination of the thresholding coefficient optimal value based on the minimum value of Shannon entropy, which is calculated for the filtered data.

The algorithm works in such a way that if the value of Shannon entropy increases at the first step of thresholding coefficient change, the filtering process is stopped. In this case the studied data do not need filtering process.

Stage III. Gene expression profiles reducing.

The aim of this stage is division of the studied gene expression profiles into informative and non-informative in terms of complex use of statistical criteria and Shannon entropy. It is assumed that if variance or average absolute value of gene expression profiles is less than the corresponding boundary values, or if Shannon entropy of the corresponding gene expression profiles is greater than the boundary value, then these profiles are non-informative and they can be removed without significant loss of useful information. The fuzzy logic system was used to determine the boundary values of the appropriate parameters. Structural block-diagram of the algorithm of gene expression profiles reducing within the framework of fuzzy logic system is presented in Fig. 4. Practical implementation of this technology involves the following steps:

1. Calculation of the variance, the average absolute value and Shannon entropy for the expression profiles of the studied genes. Formation of data in the form of corresponding vectors: $var = \{var_1, var_2,\ldots, var_m\}$, $abs = \{abs_1, abs_2,\ldots, abs_m\}$, $entr = \{entr_1, entr_2,\ldots, entr_m\}$.

2. Statistical analysis of the obtained vectors, determining the range of the appropriate parameters change.

3. Formation of the basic term-set for input variables (variance, average, Shannon entropy), and the output parameter, which determines the level of informativity of gene expression profiles *QL* (Quality).

4. Formation of the fuzzy rules, which are agreed/ with the input variables and the output parameter.

5. Determination of the boundary value of the output parameter QL_{lim}, which allows the gene expression profiles to divide into informative and non-informative. Determination of the step of the input variables changing within a given range.

6. Calculation of the output parameter *QL* for each combination of the input variables values corresponding to the appropriate gene expression profile. The result is formed as a vector: $QL = \{QL_1, QL_2,\ldots,QL_m\}$.

7. Analysis of the results. Determination the values of the input variables that correspond to the boundary value of the output parameter.

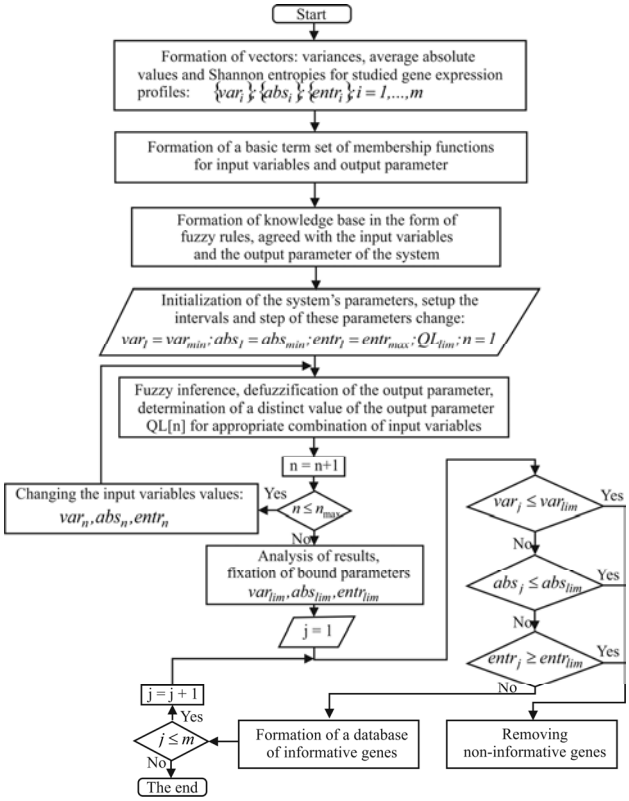

Fig. 4. Structural block-diagram of the algorithm of gene expression profiles reducing

8. A stepwise comparison of the variance, the average absolute value and Shannon entropy values of the gene expression profiles with the boundary values of the appropriate criteria. If the following conditions are true:

$$var \le var_{lim};\ abs \le abs_{lim};\ entr \le entr_{lim}$$

then this gene is allocated from the data as non-informative. Otherwise, the gene profile is recognized as informative for the further analysis.

Stage IV. Step-by-step gene expression profiles clustering within the framework of the objective clustering inductive technology.

The studies concerning development of the objective clustering inductive technology and its practical implementation based on DBSCAN and SOTA clustering algorithms are described in the papers [16-18]. The proposed methods allow us to determine objectively the parameters of the appropriate clustering algorithm operation with the use of the internal, external and complex balance clustering quality criteria. The implementation of the objective clustering inductive technology involves division of the initial dataset into two equal power subsets (containing the same quantity of pairwise similar objects). Then, the clustering process is carried out on both subsets concurrently with calculation of the internal and external clustering quality criteria at each step of the algorithm operation. At final step the complex balance criterion is calculated based on the internal and external criteria. The maximum value of the balance criterion corresponds to the optimal parameters of the appropriate clustering algorithm operation.

The use of DBSCAN clustering algorithm allows us to allocate the genes, which are identified as noise. These genes are removed from the studied data. At the second step of the clustering process the gene expression profiles are divided into two clusters with the use of SOTA clustering algorithm. These subsets are used for the following bicluster analysis.

Stage V. Bicluster analysis of the obtained subsets of the gene expression profiles.

Allocation of small groups of mutually correlated genes and samples from DNA microarray is carried out during the biclustering process. Implementation of this stage allows us to reconstruct the gene network, which will be able to reflect objectively the influence of the appropriate genes to functional possibilities of the studied biological object. Structural block-chart of biclustering technology based on "ensemble" algorithm [19] is shown in Fig. 5. Practical implementation of the technology involves the following stages:

1. The studied data preprocessing and their formation in the form of a matrix, where rows and columns are the genes and samples respectively.
2. Fixation of *simthr* parameter value, which determines the ratio of rows and columns quantity in biclusters. Setup of interval and step of thresholding coefficient value variation.
3. Data biclustering within the range of thresholding coefficient value change. Biclusters fixation at each step and calculation of the internal biclustering quality criterion.
4. Analysis of the obtained results, fixation of the thresholding coefficient value, which corresponds to the minimum of the internal quality criterion value.
5. Setup of range and step of *simthr* parameter change. Data biclustering within the given range. Fixation of biclusters at each step and calculation of the internal biclustering quality criterion.
6. Analysis of the obtained results, fixation of the ratio of rows and columns quantity in biclusters, which corresponds to the minimum of the internal quality criterion value.
7. Data biclustering with the use of the "ensemble" algorithm optimal parameters. Fixation of the biclusters.

Stage VI. Gene regulatory networks reconstruction and validation of the obtained models.

The technologies of gene regulatory networks reconstruction and validation are presented in [20]. The reconstruction of gene networks was performed based on correlation inference algorithm. The optimal topology of the obtained gene networks was determined on the basis of the maximum value of general Harrington desirability index, which contains as its components the topological parameters of networks. Validation of the obtained models was performed based on the comparison analysis of the interconnection between the appropriate genes in the basic network and in the networks based on the obtained biclusters. ROC-analysis was used to calculate the relative criterion, which indicates a quality of the obtained gene networks.

III. Results of the Simulation and Discussion

The DNA microchips of patients, which were investigated on diseases Alzheimer [21] and Parkinson [22], were used during simulation process. The first data contained 75 microchips, 16 samples were in the second database. Each of the simple contained 54675 genes. Fig. 6 shows the results of the simulation to determine the optimal combination of the methods of DNA microchip processing in order to evaluate the expression of the studied genes.

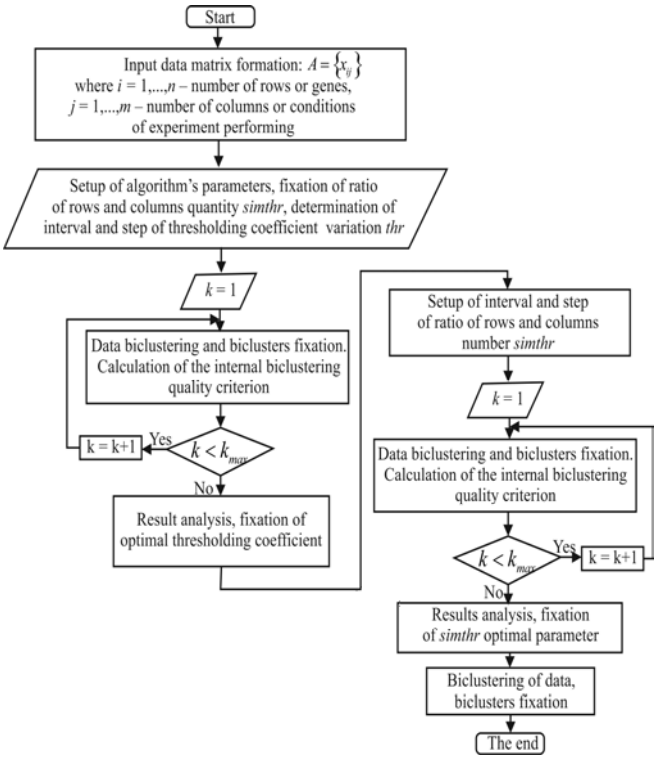

Fig. 5. Structural block-charts of biclustering technology based on "ensemble" biclustering algorithm

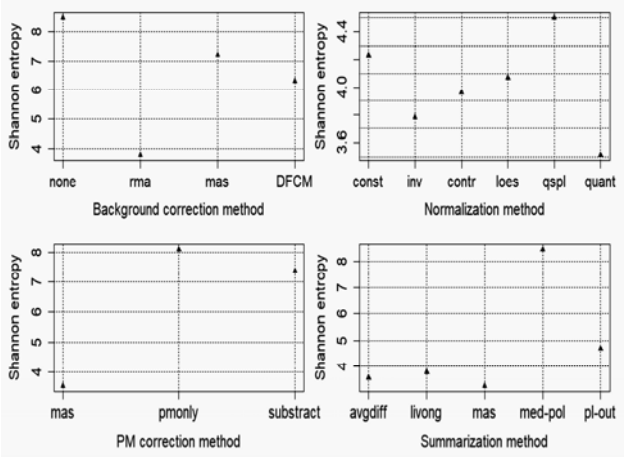

Fig. 6. Charts of distribution of Shannon entropy average values versus the used method of DNA microchip processing

Analysis of the simulation results allows us to conclude that in terms of the minimum values of Shannon Entropy the following methods are optimal for the studied DNA microchips: rma method background correction, quantile normalization, and mas methods PM correction and summarization. Fig. 7 shows the simulation results to determine the wavelet filter optimal parameters. Biorthogonal wavelet bior1.5 was used during the simulation process.

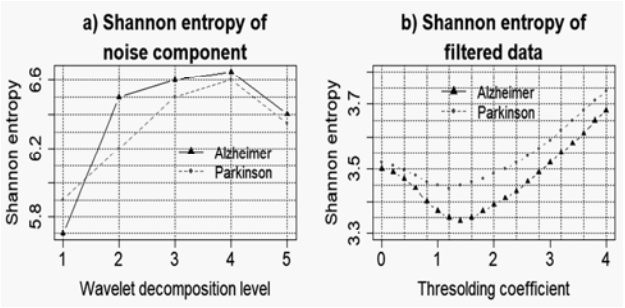

Fig. 7. Results of the simulation to determine the wavelet filter optimal parameters

The analysis of the obtained results allows concluding that in both cases the fourth wavelet decomposition level is the optimal one therefore Shannon entropy for the allocated noise component achieved the maximum values in these cases. The optimal thresholding coefficient values are 1.4 and 1.2 for gene expression profiles of patients, who were investigated on diseases Alzheimer and Parkinson respectively. In these cases, Shannon Entropies for the filtered data are minimal that indicates the maximum informativity of the studied gene expression profiles.

The next stage of data processing is the reducing of gene expression profiles based on fuzzy logic system with the use of statistical criteria and Shannon entropy. The variance, average and Shannon entropy of gene expression profiles were used as input variables. The quality of gene expression profiles was used as output parameter. The range of the output parameter change was divided into five equal intervals (very low, low, median, high, very high). The genes, which were indicated as very high by quality parameter, were allocated for the following investigation. The Gaussian and triangular membership functions were used for the input and output variables respectively. Fig. 8 and Fig. 9 show the simulation results for the determination of the input parameters boundary values.

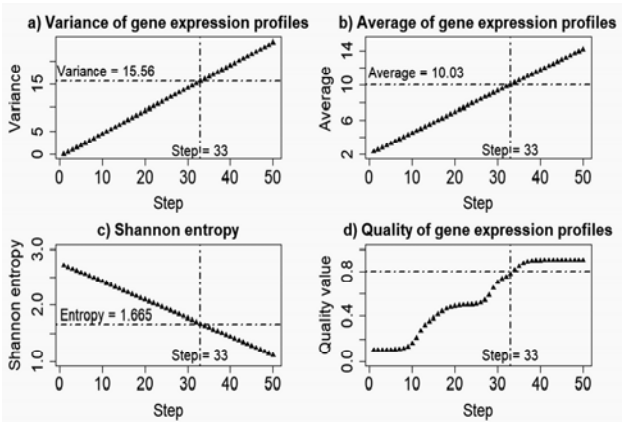

Fig. 8. Results of the simulation to determine the input parameters boundary values in the case of Alzheimer disease

As result of the simulation, the number of genes was changed from 54675 to 2037 in the case of Alzheimer disease and from 54675 to 1979 in the case of Parkinson disease. The implementation of the following cluster-bicluster technology was performed in three steps. Firstly, the gene expression profiles which were identified as noise were removed from the data with the use of DBSCAN clustering algorithm within the framework of objective clustering inductive technology. The number of genes was changed from 2037 to 1771 in the case of Alzheimer disease and from 1979 to 1649 in the case of Parkinson disease. Then, the obtained data were divided into two clusters using SOTA clustering algorithm. 770 and 1001 genes were in the first and in the second clusters respectively in the case of Alzheimer disease. Clusters in the case of Parkinson disease contained 606 and 1043 genes respectively. Finally, five biclusters were obtained from each cluster using biclustering technology, which is presented in Fig. 5.

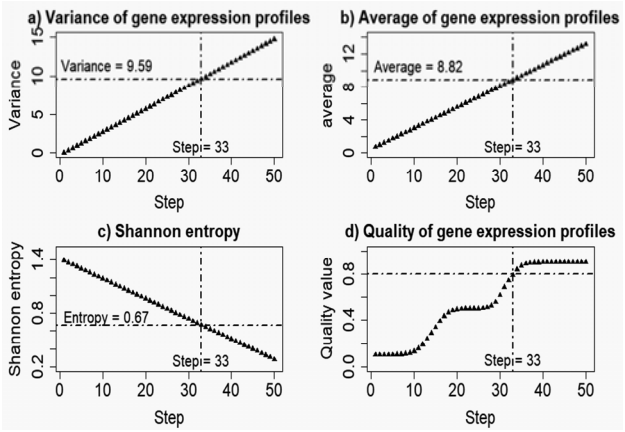

Fig. 9. Results of the simulation to determine the input parameters boundary values in the case of Parkinson disease

Gene regulatory networks reconstruction and their validation was performed with the use of the technology, which is described in detail in [20]. Fig. 10 and Fig. 11 present the results of the obtained model validation. The analysis of the obtained results allows us to conclude about the high efficiency of the proposed technology of gene expression profiles processing because the value of the relative validly criterion is high (>>1) for all obtained models. This fact indicates a high present of coincidence of appropriate genes interconnection in basic gene network and networks based on the obtained biclusters.

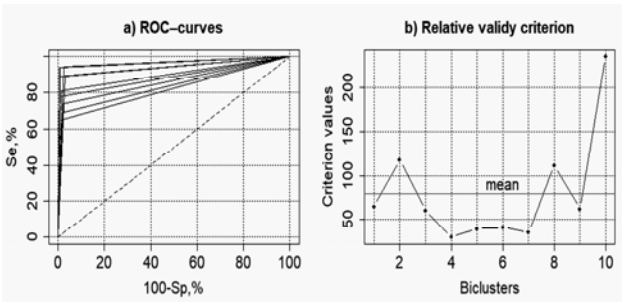

Fig. 10. Results of the gene networks validation in the case of Alzheimer disease

III. Conclusion

The results of the practical implementation of the information technology of gene expression profiles processing for the purpose of the gene regulatory networks reconstruction and validation are presented in the paper. The patients' DNA microchip data, which were investigated on

Alzheimer and Parkinson diseases were used during simulation process. The step-by-step procedure of the studied data processing included: determination of the optimal combination of the methods for evaluation of the genes expression array at the first step, determination of the wavelet filter optimal parameters and filtration the studied profiles at the second step, genes reducing, clustering, biclustering, and gene network reconstruction and validation at the last step. The obtained results have shown high efficiency of the proposed technology. The perspective of the authors' research is the implementation of the proposed technology for reconstruction of different types of gene regulatory networks.

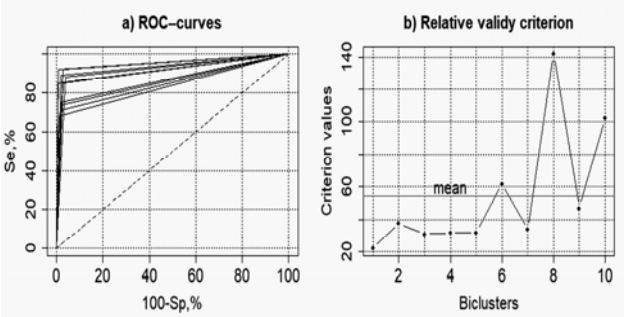

Fig. 11. Results of the gene networks validation in the case of Parkinson disease

REFERENCES

[1] D. Zak, R. Vadigepalli, E. Gonye, F. Doyle, at al. "Unconventional systems analysis problem in molecular biology: a case study in gene regulatory network modeling," Computational and Chemical Engineering, 29(3), pp. 547-563, 2005.

[2] M. Schena, and R. W. Davis. "Microarray biochip technology," Eaton Publishing, pp. 1-18, 2000.

[3] J. M. Heather, and B. Chain, "The sequence of sequencers: The history of sequencing DNA," Genomics, vol. 107, pp. 1-8, 2016.

[4] G. Setlak, Y. Bodyanskiy, I. Pliss, O. Vynokurova, D. Peleshko, and I. Kobylin, "Adaptive fuzzy clustering of multivariate short series with unevenly distributed observations based on matrix neuro-fuzzy self-organizing network," Advances in Intelligent Systems and Computing, 643, pp. 308-315, 2018.

[5] Y. Bodyanskiy, O. Vynokurova, V. Savvo, T. Tverdokhlib, and P. Mulesa, "Hybrid clustering-classification neural network in the medical diagnostics of the reactive arthritis," International Journal of Intelligent Systems and Applications, 8 (8), pp. 1-9, 2016.

[6] Y. Bodyanskiy, O. Tyshchenko, and D. S. Kopaliani, "An evolving connectionist system for data stream fuzzy clustering and its online learning," Neurocomputing, 262, pp. 41-56, 2017.

[7] Z. Hu, Y. Bodyanskiy, O. Tyshchenko, and O. Boiko, "A neuro-fuzzy Kohonen network for data stream possibilistic clustering and its online self-learning procedure," Applied soft computing, 2017.

[8] V. Mashkov, "Task allocation among aggents of restricted alliance," Eighth IASTED International Conference on Intelligent Systems and Control, ISC 2005, pp. 13-18, 2005.

[9] V. A. Mashkov, and O. V. Barabash, "Self-testing of multimodule systems on optimal check-connection structures," Engineering Simulation, 13 (3), pp. 479-492, 1996.

[10] P. Bidyuk, A. Gozhyj, I. Kalinina, V. Gozhyj. Analysis of uncertainly types for model building and forecacting dynamic processes, Advances in Intelligent Systems and Computing, 689, pp. 66-78, 2018.

[11] A. Gozhyj, I. Kalinina, and V. Gozhyj, "Fuzzy cognitive analysis and modeling of water quality," IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017, 1, art. no. 8095092, pp. 289-293, 2017.

[12] P. Bidyuk, A. Gozhyj, I. Kalinina, and V. Gozhyj, "Methods for processing uncertainties in solving dynamic planning problems," 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, art. no. 8098757, pp. 151-155, 2017.

[13] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "On biclustering of gene expression data," Current Bioinformatics, vol. 5, pp. 204-216, 2010.

[14] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," Journal of Biomedical Informatics, vol. 57 pp. 163-180, 2015.

[15] J. Hausser and K. Strimmer, "Entropy inference and the james-stein estimator with application to nonlinear gene association networks," Journal of Machine Learning Research, vol. 10, pp.1469–1484, 2009.

[16] S. Babichev, V. Lytvynenko, M. Korobchynskyi, and M. A. Taif, "Objective clustering inductive technology of gene expression sequences features," Communications in Computer and Information Science, vol. 716, pp.359-372, 2016.

[17] S. Babichev, V. Lytvynenko, J. Skvor, and J. Fiser. "Model of the objective clustering inductive technology of gene expression profiles based on SOTA and DBSCAN clustering algorithms," Advances in Intelligent Systems and Computing, vol. 689, pp. 21-39, 2018.

[18] S. Babichev, M. A. Taif, V. Lytvynenko, and V. Osypenko, "Criterial analysis of gene expression sequences to create the objective clustering inductive technology," IEEE 37th International Conference on Electronics and Nanotechnology, ELNANO 2017, pp. 244-248, 2017.

[19] S. Kaiser, Biclustering: Methods, Software and Application, Minchin, 2011.

[20] S. Babichev, M. Korobchynskyi, O. Lahodynskyi, O. Korchomnyi, and V. Borynskyi, "Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles," East-European journal of enterprise technologies, vol. 1/4 (91), pp. 19-32, 2018.

[21] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, et al. "Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons," Proc. Nat. Acad. Sci. USA, vol. 105(11), pp. 4441-4446, 2008.

[22] B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, et al. "PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease," Sci. Transl. Med., vol. 2(52), pp. 52-73, 2010.