I.J. Intelligent Systems and Applications, **2018, 7, 1-10**
Published Online July 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2018.07.01

An Effectiveness Evaluation of Information Technology of Gene Expression Profiles Processing for Gene Networks Reconstruction

Sergii Babichev
Jan Evangelista Purkyně University in Ustí nad Labem, Ustí nad Labem, Czech Republic
E-mail: sergii.babichev@ujep.cz

Maksym Korobchynskyi, Serhii Mieshkov, Oleksandr Korchomnyi
Military-Diplomatic Academy named after Eugene Bereznyak, Kyiv, Ukraine
E-mail: maks_kor@ukr.net

Received: 03 March 2018; Accepted: 24 May 2018; Published: 08 July 2018

***Abstract*—**The paper presents the research results concerning an effectiveness evaluation of information technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction. The information technology is presented as a structural block-chart of step-by-step stages of the studied data processing. The DNA microchips of patients, who were investigated on different types of cancer, were used as experimental data. The optimal parameters of data processing algorithm at appropriate stage of this process implementation by quantity criteria of data processing quality were determined during simulation. Validation of the reconstructed gene networks was performed with the use of ROC-analysis by comparison of character of genes interconnection in both the basic network and networks reconstructed based on the obtained biclusters.

***Index Terms*—**Gene expression profiles, Filtering, Reducing, Clustering, Biclustering, Gene network reconstruction, Gene network validation.

I. Introduction

Reconstruction and simulation of gene regulatory networks (GRN) based on gene expression profiles is one of the current problems of modern bioinformatics. The implementation of this process corresponds to better understanding of genes interactions character and influence of these interactions to functional possibilities of biological objects. The gene expression profile is a vector of gene expressions, which are determined on different conditions of experiment performing. Two technologies of gene expressions array formation are actual nowadays: DNA microchip technology [1,2] and mRNA sequencing method [3,4]. Each of them has own advantages and disadvantages. However, in any case, the final result is a matrix of genes expressions. The rows of this matrix are genes and the columns present the conditions of experiment performing. Peculiarities of the experimental data are high dimension of feature space and existence of complex noise component, which arises in the case of DNA microchip technology implementation at the stage of experiment performing and initial data formation. The complexity of GRN reconstruction is determined by the following:

- the experimental data do not allow us usually to define the network structure and character of genes interaction;
- the large number of genes, which determine the structure and amount of the network, complicates the process of the obtained results interpretation.

Solving this problem is possible by developing modern methods for processing gene expression profiles obtained by DNA microchip experiments or mRNA molecules sequencing method to reduce non-informative genes and further grouping of remaining genes using modern clustering and biclustering technologies. The final step of this process is the reconstruction and validation of gene networks and following simulation in order to evaluate their adequacy.

The first works concerning gene regulatory networks reconstruction were published at the end of the last century. The papers [5,6] presents the technologies of GRN reconstruction based on linear modeling methods and mutually correlation of gene expression profiles appropriately. The methods of GRN reconstruction and simulation based on Bayes networks and differential equation were analysed in [7,8]. However, it should be noted that the presented methods are based on the use of a small number of genes. The methods of forming an experimental data in order to reconstruct GRN were not consider in these papers. The comparison analysis of different methods of GRN reconstruction with allocation their advantages and disadvantages are presented in reviews [9-12]. However, the methods of optimizing the parameters of the proposed models and algorithms were not considered in these works.

Modern technologies for obtaining gene expression data tends to cover the maximum number of system variables [13]. For example, the technology of DNA microchip or mRNA molecules sequencing method allow measuring the expression of tens of thousands of genes concurrently. In this case, each of the studied object is characterized by a numerical vector of genes expression in the length of tens of thousands of units. GRN reconstruction based on a complete set of genes is problematic because this process requires a large amount of computer resources and the interpretation of the model is a very complicated problem. Biclustering technology is actual to group the genes and samples nowadays [14-16]. However, it should be noted that the application of this technology does not solve the problem of studied data grouping. The use of biclustering algorithms allows us to obtain clusters of mutually correlated genes and samples, but there is a problem of the choice of biclusters quantity and the level of detail of this process. The results of comparison analysis of biclustering algorithms effectiveness with the use of tested data and gene expression profiles are presented in [17]. The hybrid model of gene expression profiles cluster-bicluster analysis was proposed as the result of research. Data clustering within the framework of this model was carried out with the use of objective clustering inductive technology, the main conception of which is described in [18]. The paper [19] presents the results of research concerning practical implementation the objective clustering inductive technology based on SOTA (Self Organizing Tree Algorithm) and DBSCAN (Density Based Spatial Clustering of Application with Noise) clustering algorithms. The gene expression profiles were used as the test data during simulation process. The results of research concerning development of technology for reconstruction and validation of gene networks based on the obtained biclusters are presented in [20]. However, it should be noted that in spite of achievements in this subject area there are some unsolved problems. So, nowadays there are no effective technologies for both gene expression profiles preprocessing and following grouping genes and samples for purpose of gene regulatory network reconstruction. This fact indicates the relevance of the presented research topic.

The aim of the research is an effectiveness evaluation of the information technology of gene expression profiles processing for gene regulatory networks reconstruction with the use of gene expression profiles of patients, who were investigated on different types of cancer.

The paper is organized as follows. Section 2 presents the structural block-chart of the information technology of gene expression profiles processing and step-by-step procedure of its implementation. Section 3 is devoted to practical implementation of the technology to process the DNA microchip data of patients, who were investigated on different type of cancer. The results of the reconstructed models of gene regulatory networks validation are presented in section 4. This section contains also the discussion of the obtained results. The conclusions are presented in section 5.

II. Information Technology of Gene Expression Profiles Processing

The architecture of the information technology of gene expression profiles processing for gene regulatory networks reconstruction and validation is presented in Fig. 1.

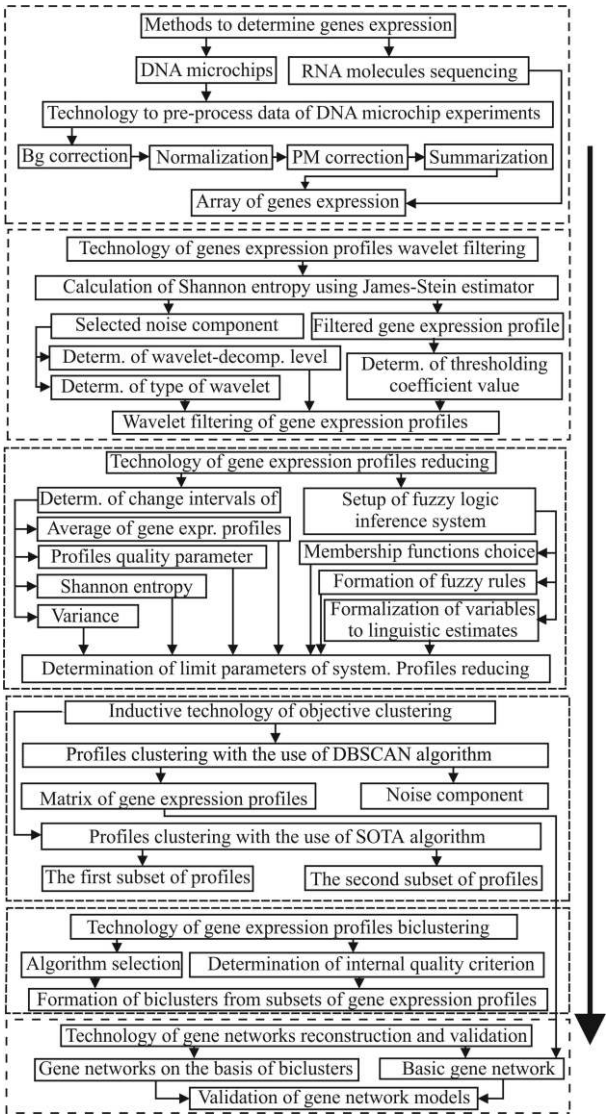


Fig.1. Information technology of gene expression profiles processing

The practical implementation of this technology involves the following stages:

Stage I. *Formation of genes expressions array*

1. Transformation of the matrix of DNA microchips light intensities to the expressions of the corresponding genes. Implementation of this stage involves four steps: background correction, normalization, PM correction and summarization. Each of these steps can be implemented with the use of different methods. Choice of the optimal combination of the methods within the framework of the

proposed technology was performed based on minimum value of Shannon entropy calculated with the use of James-Stein shrinkage estimator [21].

Stage II. *Gene expression profiles wavelet filtering*

2. Determination of the wavelet filter optimal parameters (type of wavelet, wavelet decomposition level and thresholding coefficient value).
3. Step-by-step gene expression profiles wavelet filtering in accordance with the technology presented in [22].

Stage III. *Gene expression profiles wavelet reducing*

4. Calculation of variance, average absolute value and Shannon entropy for the studied gene expression profiles. Statistical analysis of the obtained vectors. Setup of range of the appropriate parameters change. It was supposed that if values of variance and average absolute value of gene expression profile is less and Shannon entropy is more of the appropriate boundary values, this gene is removed from data as non-informative. Determination of the input parameters boundary values was performed with the use of fuzzy logic inference technology [23-26].
5. Setup of fuzzy logic inference system. Formation of the basic term-set for input variables (variance, average, Shannon entropy), and output parameter, which determines the level of informativity of gene expression profiles *QL* (Quality). Formation of fuzzy rules, which are agreed between both the input variables and output parameter.
6. Determination of the boundary value of the output parameter QL_{lim}, which allows us the gene expression profiles to divide into informative and non-informative. Determination of the step of the input variables changing within a given range.
7. Determination of the input parameters boundary values according to technology presented in [27]. Gene expression profiles reducing. Formation of a new array of gene expression profiles.

Stage IV. *Gene expression profiles clustering*

8. Formation of two equal power gene expression profiles subsets (contain the same quantity of pairwise similar profiles).
9. Determination of optimal parameters of clustering algorithms SOTA and DBSCAN using inductive technology of objective clustering [18,19].
10. Gene expression profiles clustering with the use of DBSCAN clustering algorithm. Allocated genes, which are identified as noise. Formation of a new data array.
11. Gene expression profiles clustering with the use of SOTA clustering algorithm. Formation of two clusters of gene expression profiles for the following step of bicluster analysis.

Stage V. *Gene expression profiles biclustering*

12. Determination of "ensemble" biclustering algorithm optimal parameters within the framework of the technology presented in [17].
13. Gene expression profiles biclustering. Allocation of groups of mutually correlated genes and samples for the following step of gene regulatory networks reconstruction and validation.

Stage VI. *Gene regulatory networks reconstruction with the use of correlation inference algorithm*

14. Determination of thresholding coefficient optimal parameters for the basic gene network (reconstructed based on the set of gene expression profiles obtained at step 10 of this procedure) and for gene networks reconstructed on the basis of the obtained biclusters according to the technology presented in [20].
15. Reconstruction of gene regulatory networks.

Stage VII. *Validation of the reconstructed models of gene networks*

16. ROC-analysis of the obtained models of gene networks according to technology presented in [20]. Calculation of the validation relative criteria for obtained models of gene networks.
17. Analysis of the obtained results.

III. Practical Implementation of the Information Technology of Gene Expression Profiles Processing

Three types of DNA microchips of patients, who were investigated on different types of cancer, were used during simulation process. The first data contained 64 DNA microchips investigated on colorectal cancer [28]. 32 patients from them were healthy and other 32 patients had this type of disease. The second data included 26 DNA microchips investigated on prostate cancer [29]. 13 patients from them were healthy and other 9 ones were ill. The third data included 88 DNA microchips investigated on lung cancer [30]. 44 patients from them were recognized as healthy. Fig. 2 shows the results of the first stage of the information technology implementation. This figure contains the charts of Shannon entropy values distribution versus the used method and step of the DNA microchip data processing.

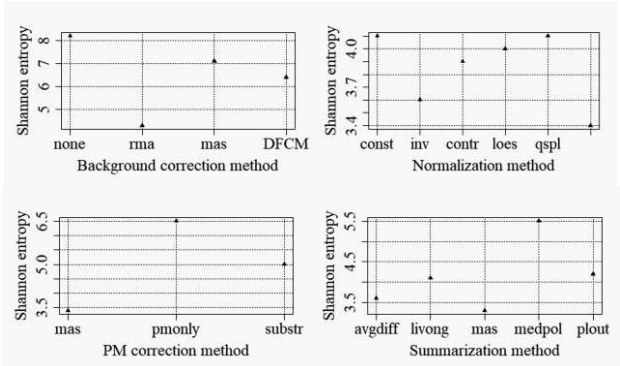


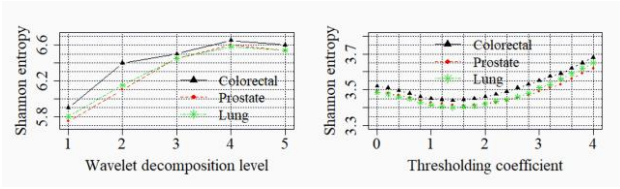
Fig.2. Charts of the average of Shannon entropy values distribution versus the step and method of the DNA microchip data processing

Analysis of the obtained results allows us to conclude that optimal in terms of Shannon entropy criterion is the following combination of the methods: "rma"

background correction method, “quantile” normalization method and “mas” methods of PM correction and summarization. The average of Shannon entropy in this case is the minimal that indicates the maximum informativity of the obtained gene expression profiles to compare with the use of other combination of the methods. The obtained data are presented as a matrix, where rows are the studied samples or condition of the experiment performing and columns are the studied genes. The sizes of the studied data are the following:

- patients’ array investigated on colorectal cancer: (64×54675);
- patients’ array investigated on prostate cancer: (26×22277);
- patients’ array investigated on lung cancer: (88×54675).

According to the proposed technology, the next stage of the data processing is the obtained gene expression profiles wavelet filtering. Determination of optimal parameters of wavelet filter was performed by the method, which was described in detail in [22]. Biorthogonal wavelet *bior1.5* was used during simulation process. The optimal level of wavelet decomposition was determined based on maximum value of Shannon entropy for allocated noise component. Determination of thresholding coefficient optimal value was carried out on the basis of minimum value of Shannon entropy for filtered data. The results of the simulation are presented in Fig. 3.



a b

Fig.3. Results of the simulation to determine the wavelet filter optimal parameters: a) chart of Shannon entropy of allocated noise component versus the wavelet decomposition level; b) chart of Shannon entropy of filtered data versus the thresholding coefficient

The analysis of the obtained results allows us to conclude that in all cases the fourth wavelet decomposition level is the optimal one therefore the value of Shannon entropy in these cases for the allocated noise component achieved the maximum values. The optimal thresholding coefficient values are 1.4 for patients’ gene expression profiles investigated on colorectal cancer and lung cancer. In the case of the DNA microarray of patients, who were investigated on prostate cancer the thresholding coefficient optimal value was taken 1.6. In these cases Shannon entropy criterion for the filtered data was minimal. This fact indicates the maximum informativity of the studied gene expression profiles.

Gene expression profiles reducing involves removing genes whose expression profiles have variance and average absolute value less and Shannon entropy more than corresponding boundary values. The technology of gene expression profiles reducing based on fuzzy logic inference system with the use of statistical criteria and Shannon entropy has been proposed in [27]. Variance, average absolute value 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 sections (very low, low, median, high and very high). Statistical characteristics of input variables for the studied gene expression profiles are presented in tables 1-3. Formalization of the input variables to linguistic estimates for the studied data within the framework of the proposed fuzzy logic inference model are presented in tables 4-6. The Gaussian and triangular membership functions were used for the input and output variables respectively. The genes, which were indicated as very high by quality parameter (≥0.8) for patients’ data, who were studied on colorectal cancer and lung cancer and as high quality (≥0.6) for patients’ data investigated on prostate cancer, were allocated for the following investigation.

Table 1. Variation of gene expression profiles variance

Type of cancer	Min	Median	Mean	Max
Colorectal	0.004	0.07	0.143	12.68
Prostate	0.002	0.097	0.21	7.8
Lung	0.007	0.11	0.31	16.6

Table 2. Variation of gene expression profiles average absolute values

Type of cancer	Min	Median	Mean	Max
Colorectal	2.63	4.87	5.41	14.49
Prostate	3.68	7.34	7.44	14.38
Lung	2.89	4.62	5.37	14.24

Table 3. Variation of gene expression profiles Shannon entropy

Type of cancer	Min	Median	Mean	Max
Colorectal	0.67	2.63	2.60	2.77
Prostate	0.19	1.74	1.68	1.792
Lung	0.45	2.88	2.83	3.09

Table 4. Formalization of input parameters to linguistic estimates for patients’ data investigated on colorectal cancer

Input parameters	Range	Terms of linguistic estimate
Variance (Vr)	0–13	mean=2, sd=1.2 – «Low» mean=5.5, sd=1.2 – «Md» mean=11.5, sd=1.6 – «Hg»
Avereage absolute value (Abs)	2–15	mean=3.5, sd=1.2 – «Low» mean=7, sd=1.2 – «Md» mean=12, sd=1.6 – «Hg»
Shannon entropy (Entr)	0.6–2.8	mean=1, sd=0.2 – «Low» mean=1.8, sd=0.15 – «Md» mean=2.4, sd=0.15 – «Hg»

Table 5. Formalization of input parameters to linguistic estimates for patients' data investigated on prostate cancer

Input parameters	Range	Terms of linguistic estimate
Variance (Vr)	0–8	mean=1, sd=0.6 – «Low» mean=3.4, sd=0.6 – «Md» mean=7, sd=1 – «Hg»
Avereage absolute value (Abs)	3–15	mean=4, sd=1 – «Low» mean=7, sd=1 – «Md» mean=13, sd=1.4 – «Hg»
Shannon entropy (Entr)	0.18–1.8	mean=0.4, sd=0.2 – «Low» mean=1.1, sd=0.15 – «Md» mean=1.6, sd=0.15 – «Hg»

Table 6. Formalization of input parameters to linguistic estimates for patients' data investigated on lung cancer

Input parameters	Range	Terms of linguistic estimate
Variance (Vr)	0–17	mean=3, sd=1.5 – «Low» mean=7, sd=1.5 – «Md» mean=14, sd=2 – «Hg»
Avereage absolute value (Abs)	2.8–14.3	mean=3.5, sd=1.2 – «Low» mean=7, sd=1.2 – «Md» mean=12, sd=1.6 – «Hg»
Shannon entropy (Entr)	0.4–3.1	mean=1, sd=0.2 – «Low» mean=1.9, sd=0.15 – «Md» mean=2.6, sd=0.15 – «Hg»

Fig. 4 shows the simulation results for the determination of the input parameters boundary values for patients' data investigated on colorectal cancer. The same results were obtained for other studied data. The range of change of the input parameters was divided into 50 equal sections. The variance and average absolute values were changed from minimum to maximum and Shannon entropy was changed from maximum to minimum during simulation process.

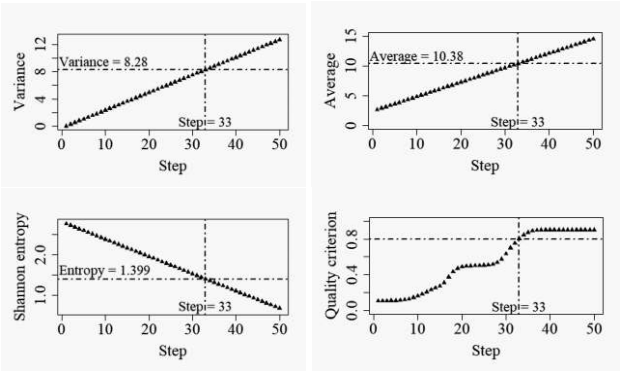


Fig.4. Results of the simulation to determine the input parameters boundary values in the case of colorectal cancer data using

As the result of the simulation the initial data were transformed as following:

- patients' gene expression profiles investigated on colorectal cancer: $(64 \times 54675) \rightarrow (64 \times 1758)$;
- patients' gene expression profiles investigated on prostate cancer: $(26 \times 22277) \rightarrow (26 \times 1990)$;
- patients' gene expression profiles investigated on lung cancer: $(88 \times 54675) \rightarrow (88 \times 2126)$.

The next stage of the proposed technology implementation involves cluster-bicluster analysis in order to allocate groups of genes and samples for following gene networks reconstruction. Structural block-chart of cluster-bicluster technology based on DBSCAN and SOTA clustering algorithms and biclustering method "ensemble" is presented in Fig. 5 [17-19]. Practical implementation of this technology involves the following steps:

1. Division of the initial gene expression profiles dataset into two equal power subsets, which contain the same quantity of pairwise similar objects;
2. Determination of the optimal parameters for DBSCAN and SOTA clustering algorithms operation according to the technology presented in [19].
3. Data clustering with the use of DBSCAN algorithm. Allocation of a noise component. Formation of a new matrix of gene expression profiles for the following processing.
4. Clustering of the obtained gene expression profiles with the use of SOTA clustering algorithm. Formation of two subsets of gene expression profiles for the following bicluster analysis.
5. Determination of the optimal parameters for "ensemble" biclustering algorithm according to the technology presented in [17].
6. Gene expression profiles biclustering. Formation of biclusters, which contain mutually correlated genes and samples.

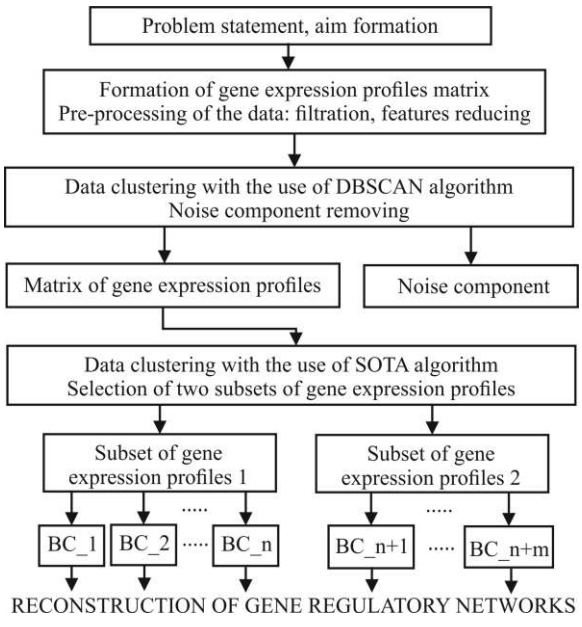


Fig.5. Structural block-chart of cluster-bicluster technology

Fig. 6 shows the simulation results to determine the optimal parameters of DBSCAN clustering algorithm operation within the framework of the objective clustering inductive technology. Minimal quantity of points (MinPts) inside epsilon-neighborhood (EPS) was changed from 3 to 4. Increase of MinPts value promoted to unsatisfactory results because many little clusters and many genes, which were identified as noise, appeared in these cases. The analysis of the obtained results allows us

to determine the following parameters of DBSCAN clustering algorithm operation: gene expression profiles of patients, who were investigated on colorectal cancer: 1) EPS = 0.22, MinPts = 3; 2) EPS = 0.27, MinPts = 4. In the first case 92 genes and in the second 343 ones were identified as noise. The second case was selected. The number of genes was change from 1758 to 1666;

- gene expression profiles of patients, who were investigated on prostate cancer: 1) EPS = 0.19, MinPts = 3; 2) EPS = 0.26, MinPts = 4. In the first case 285 genes and in the second 428 ones were identified as noise. The first case was selected. The number of genes was change from 1990 to 1705;
- gene expression profiles of patients, who were investigated on lung cancer: 1) EPS = 0.2, MinPts = 3; 2) EPS = 0.42, MinPts = 3; 3) EPS = 0.2, MinPts = 4; 4) EPS = 0.4, MinPts = 4. In the first case 523 genes, in the second 331, in the third 570 and in the fourth 470 genes were identified as noise. The second case was selected. The number of genes was change from 2126 to 1795.

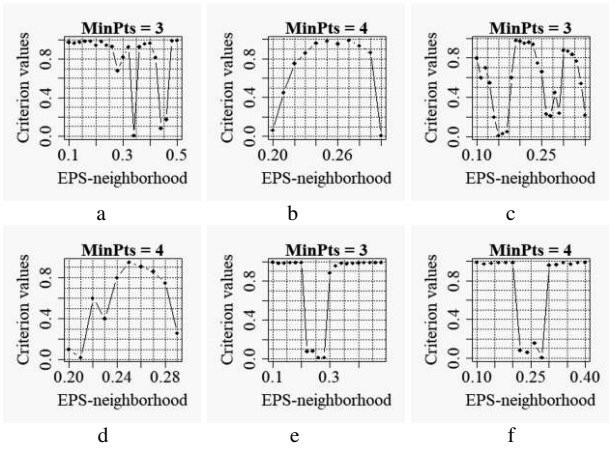


Fig.6. Charts of complex balance criterion versus the EPS and MinPts values for patients' gene expression profiles, who were investigated on: a,b) colorectal cancer; c,d) prostate cancer; e,f) lung cancer

The results of the DBSCAN clustering algorithm operation is shown in Fig. 7.

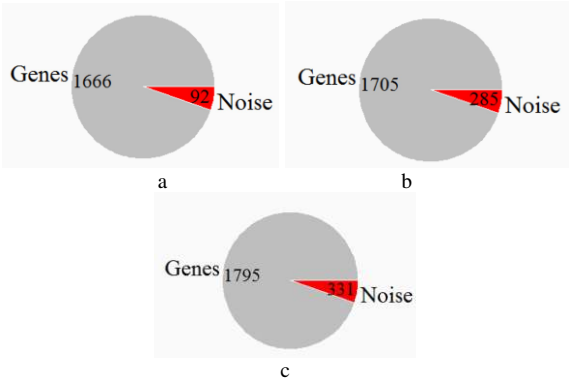


Fig.7. Results of DBSCAN clustering algorithm operation to allocate the genes, which were identified as noise in the case of a) colorectal cancer data; b) prostate cancer data; c) lung cancer data

Fig. 8 shows the results of simulation to determine the scell optimal parameter of SOTA clustering algorithm. The value of this parameter was changed within the range from 0.001 to 0.2 with step 0.005. The general Harrington desirability index (complex balance criterion) was calculated at each step of data clustering according to the technology presented in [19]. The following values of scell parameter were determined during simulation process:

- gene expression profiles of patients, who were investigated on colorectal cancer: scell = 0.051;
- gene expression profiles of patients, who were investigated on prostate cancer: scell = 0.051;
- gene expression profiles of patients, who were investigated on lung cancer: 0.026.

In all cases the studied gene expression profiles were divided into two clusters. The results of the algorithm operation is presented in Fig. 9.

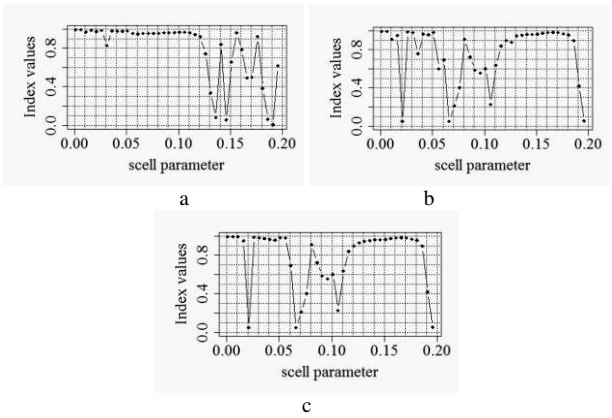


Fig.8. Charts of complex balance criterion versus the scell value for gene expression profiles of patients, who were investigated on: a) colorectal cancer; b) prostate cancer; c) lung cancer

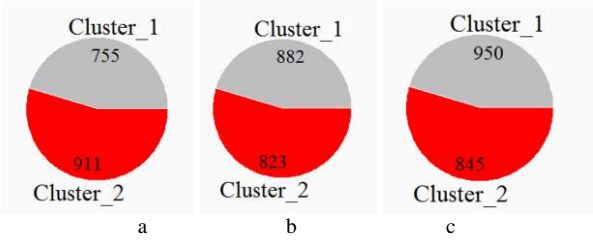


Fig.9. Results of SOTA clustering algorithm operation to divide the gene expression profiles into two clusters in the case of a) colorectal cancer data; b) prostate cancer data; c) lung cancer data

Bicluster analysis was performed according to the technology presented in [17] with the use of "ensemble" biclustering algorithm. Firstly, the optimal parameters of "ensemble" biclustering algorithm were determined. Then, the biclustering of the studied gene expression profiles was carried out with the use of the optimal parameters of the biclustering algorithm operation. Five the largest biclusters were selected from each subset of the studied gene expression profiles. The results of biclustering for all types of gene expression profiles are presented in tables 7-9.

Table 7. Results of biclustering of patients' gene expression profiles investigated on colorectal cancer

Objects		911 of genes				
Biclust.		BC1	BC2	BC4	BC5	BC8
Genes		108	97	18	24	243
Samples		29	26	25	23	32
Objects		755 of genes				
Biclust.		BC1	BC2	BC3	BC4	BC8
Genes		118	74	107	161	36
Samples		30	29	28	32	17

Table 8. Results of biclustering of patients' gene expression profiles investigated on prostate cancer

Objects	823 of genes				
Biclust.	BC1	BC2	BC4	BC5	BC6
Genes	79	84	97	72	126
Samples	8	5	4	7	4
Objects	882 of genes				
Biclust.	BC1	BC2	BC3	BC4	BC5
Genes	42	54	58	42	76
Samples	7	6	5	4	3

Table 9. Results of biclustering of patients' gene expression profiles investigated on lung cancer

Objects	950 of genes				
Biclust.	BC1	BC3	BC4	BC5	BC6
Genes	54	29	19	27	16
Samples	29	29	28	28	14
Objects	845 of genes				
Biclust.	BC1	BC2	BC5	BC6	BC7
Genes	18	41	17	21	18
Samples	26	30	52	29	26

IV. Gene Regulatory Networks Reconstruction and Validation

Gene regulatory networks reconstruction and their validation were carried out with the use of technology, which was described in [20]. Correlation inference algorithm of gene regulatory network reconstruction was used during the simulation process. Firstly, the thresholding coefficient optimal value of the used algorithm was determined. At this stage it were investigated both the basic gene networks, which were reconstructed based on the data obtained as the result of DBSCAN clustering algorithm operation and gene networks, which were reconstructed based on the obtained biclusters. Gene network validation involves the comparison analysis of interconnection character between genes in both the basic networks and networks reconstructed on the basis of obtained biclusters. The technology of ROC-analysis was used at this stage. A structural block chart of gene networks validation technology is presented in Fig. 10. Practical implementation of this technology involves the following steps:

1. Data preprocessing: filtering, reducing, clustering and biclustering of gene expression profiles.
2. Reconstruction of basic gene network based on the data from the obtained clusters after BDSCAN clustering algorithm operation.
3. Reconstruction of gene networks based on the data from the relevant biclusters.
4. Determination of the quality parameters for the classification of interconnection character between genes in the appropriate networks (*TP(true positive), TN(true negative), FP(false positive), FN(false negative)*).
5. Calculation of relative quality parameters for the model quality estimation *Sc(sensitivity), Sp(specificity), FPR(false positive ratio)*.
6. Calculation of relative validation criterion: *RVC = SC/FPR.* Analysis of the obtained results.

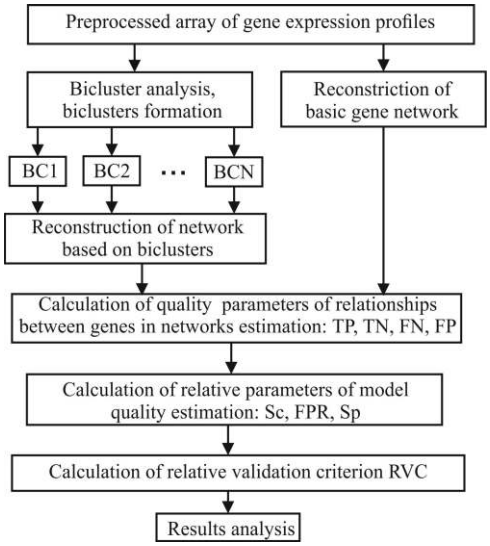


Fig.10. Structural block chart of technology of gene regulatory networks validation

Fig. 11 shows the results of validation of the reconstructed gene networks models.

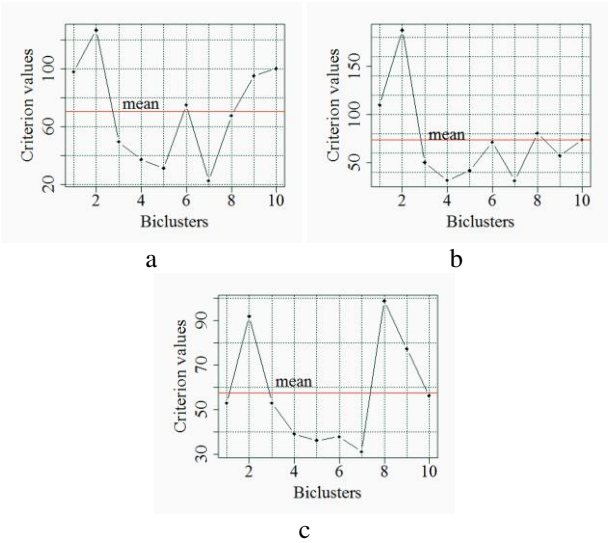


Fig.11. Results of validation of the gene networks models, which were reconstructed based on data of patients, who were studied on a) colorectal cancer; b) prostate cancer; c) lung cancer

The analysis of the obtained results allows us to conclude that the reconstructed gene network have high level of adequacy since the value of relative validation criterion in all cases is essentially more than 1(statistically insignificant). The gene networks reconstructed on the basis of data of the biclusters for which the value of the relative validation criterion is greatest have the particular interest for further modeling. These networks are the most adequate to the basic gene network reconstructed based on full dataset of studied genes and samples. Further modeling of gene regulatory networks allows us to consider the particularities of influence of the expression of individual genes to the expression values of other genes of network for purpose of better understanding character of interconnection between genes for different type of diseases and for different state of the investigated biology object.

V. Conclusion

The research concerning practical implementation of the information technology of gene expression profiles processing for purpose of gene regulatory networks reconstruction and validation has been presented in the paper. The DNA microchips of patients, who were investigated on colorectal cancer, prostate cancer and lung cancer, have been used during simulation process. The first step involved determination of optimal combination methods to obtain the array of genes expression profiles. Shannon entropy calculated with the use of James-Stein shrinkage estimator has been used as criterion during implementation of this stage. The following combination of methods has been determined as the result of simulation: “rma” background correction method, “quantile” normalization method and “mas” methods of PM correction and summarization.

The second stage involved wavelet filtering of the studied gene expression profiles. Biorthogonal wavelet bior1.5 have been used during simulation process. The optimal level of wavelet decomposition was determined based on maximum value of Shannon entropy for allocated noise component. Determination of thresholding coefficient optimal value was carried out on the basis of minimum value of Shannon entropy for filtered data. The results of the simulation have shown that in all cases the fourth level of wavelet decomposition was an optimal. The thresholding coefficients values to process the detail coefficients were the following: 1.4 for patients’ gene expression profiles, who were investigated on colorectal cancer, and 1.6 for patients’ gene expression profiles, who were investigated on prostate cancer and lung cancer. Implementation of gene expression profiles reducing technology has been performed with the use of statistical criteria and Shannon entropy. It was supposed that if values of variance and average of absolute value of gene expression profile is less and Shannon entropy is more of boundary values of the appropriate parameters, this gene is removed from data as non-informative. Determination of the input parameters boundary values has been performed with the use of fuzzy logic inference technology. The quantity of genes for the following investigation has been essentially decreased during simulation process.

Implementation of the cluster-bicluster technology involved three steps. Firstly, the genes which were identified as noise have been removed by DBSCAN clustering algorithm operation. Then, the gene expression profiles have been divided into two groups with the use of SOTA clustering algorithm. Finally, the biclusters have been allocated by the use of “ensemble” biclustering algorithm. Determination of optimal parameters of the clustering algorithms has been carried out within the framework of objective clustering inductive technology. The optimal parameters of biclustering algorithm have been performed based on minimum value of internal biclustering quality criterion.

The correlation inference algorithm has been used for gene networks reconstruction. Topology of the reconstructed networks is determined by thresholding coefficient value. Determination of optimal value of the thresholding coefficient has been performed on the basis of maximum value of the complex topological parameter. The finally stage of the proposed technology implementation is validation of the reconstructed networks. ROC-analysis has been used to implement of this stage. The charts of relative validation criteria distribution for reconstructed gene networks have been constructed. The analysis of the obtained results has shown high effectiveness of the proposed technology because the value of relative validation criterion for all gene networks was essentially more than 1. This fact indicates the high level of adequate of networks reconstructed based on biclusters to network reconstructed on the basis of full set of genes.

References

[1] X. Zhang, Z. Zhou, Y. Jiao, Y. Niu and Y. Wang, “A visual cryptography scheme-based DNA microarrays”, International Journal of Performability Engineering, vol. 14(2), pp. 334–340, 2018.

[2] Shukla, S., Agarwal, A.K., Lakhmani, A., “MICROCHIPS: A leading innovation in medicine”, Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom, art. no. 7724256, pp. 205-210, 2016.

[3] X. Wu, B. Yang, I. Udo-Inyang, S. Ji, D. Ozog, L. Zhou and Q.-S. Mi, “Research Techniques Made Simple: Single-Cell RNA Sequencing and its Applications in Dermatology”, Journal of Investigative Dermatology, vol. 138(5), pp. 1004–1009, 2018.

[4] Wang, L.Y., Guo, J., Cao, W., Zhang, M., He, J., Li, Z., “Integrated sequencing of exome and mRNA of large-sized single cells”, Scientific Reports, vol. 8 (1), art. no. 384, 2018.

[5] P. D‘haeseleer, X. Wen, S. Fuhrman, R. Somogyi, “Linear modeling of mRNA expression levels during CNS development and injury”, Pacific Symposium on Biocomputing, pp. 41–52, 1999.

[6] S. Liang, S. Fuhrman, R. Somogyi, R. Reveal, “A general reverse engineering algorithm for inference of genetic network architectures”, Pacific Symposium on Biocomputing, pp. 18–29, 1998.

[7] N. Friedman, M. Linial, I. Nachman, D. Pe'er, "Using Bayesian networks to analyse expression data", Journal of Computational Biology, 7(3–4), pp. 601–620, 2000.
[8] T. Chen, H.L. He, G.M. Church, "Modeling gene expression with differential equations", Proceedings of the Pacific Symposium on Biocomputing, pp. 29–40, 1999.
[9] K.-C. Wong, Y. Li, Z. Zhang, "Unsupervised learning in genome informatics", Unsupervised Learning Algorithms, pp. 405–448, 2016.
[10] F. Emmert-Streib, M. Dehmer, B. Haibe-Kains, "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks", Frontiers in cell and developmental biology, pp. 1–7, 2014.
[11] Zheng, G., Huang, T., "The reconstruction and analysis of gene regulatory networks", Methods in Molecular Biology, vol. 1754, pp. 137-154, 2018.
[12] Wu, H.C., Zhang, L., Chan, S.C., "Reconstruction of gene regulatory networks from short time series high throughput data: Review and new findings", International Conference on Digital Signal Processing (DSP), no. 6900761, pp. 733-738, 2014.
[13] K. Wang, L. Zhang, X. Liu, "A review of gene and isoform expression analysis across multiple experimental platforms", Chinese Journal of Biomedical Engineering, 36(2), pp. 211–218, 2017.
[14] B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, "Biclustering on expression data: A review", Journal of Biomedical Informatics, vol. 57, pp. 163–180, 2015.
[15] Rocha, O., Mendes, R., "JBiclustGE: Java API with unified biclustering algorithms for gene expression data analysis", Knowledge-Based Systems, vol. 155, pp. 83-87, 2018.
[16] Puleo, G.J., Milenkovic, O., "Correlation Clustering and Biclustering with Locally Bounded Errors", IEEE Transactions on Information Theory, 64(6), pp. 4105-4119, 2018.
[17] S. Babichev, V. Lytvynenko, V. Osypenko, M. Korobchinskyi, M. Voronenko, "Comparison analysis of biclustering algorithms with the use of artificial data and gene expression profiles", Proceedings of 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), Kiev, pp. 298–304, 2018.
[18] S. Babichev, M.A. Taif, V. Lytvynenko, M. Korobchynskyi, "Objective clustering inductive technology of gene expression sequences features", Communication in Computer and Information Science, vol. 716, pp. 359–372, 2017.
[19] S. Babichev, V. Lytvynenko, J. Skvor, 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.
[20] S. Babichev, M. Korobchynskyi, O. Lahodynskyi, O. Korchomnyi, B. Basanets, V. Borynskyi, "Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles", Eastern European Journal of Enterprise Technologies, vol. 1(4-91), pp. 19-32, 2018.
[21] 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.
[22] S. Babichev, J. Škvor, J. Fišer, V. Lytvynenko, "Technology of gene expression profiles filtering based on wavelet analysis", International Journal of Intelligent Systems and Applications, vol. 10(4), pp. 1–7, 2018.
[23] Ye.V. Bodyanskiy, A.K. Tyshchenko, A.A. Deineko, "An evolving radial basis neural network with adaptive learning of its parameters and architecture", Automatic Control and Computer Sciences, vol. 49, iss. 5, pp. 255-260, 2015.
[24] Zh. Hu, Ye.V. Bodyanskiy, O.K. Tyshchenko, "A deep cascade neuro-fuzzy system for high-dimensional online fuzzy clustering",Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing (DSMP 2016), Lviv, Ukraine, pp. 318-322, 2016.
[25] Zh. Hu, Ye.V. Bodyanskiy, O.K. Tyshchenko, "A cascade deep neuro-fuzzy system for high-dimensional online possibilistic fuzzy clustering", Proceedings of the 11th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2016, Lviv, Ukraine, pp. 119-122, 2016.
[26] Ye.V. Bodyanskiy, O.K. Tyshchenko, D.S. Kopaliani, "An Evolving Connectionist System for Data Stream Fuzzy Clustering and Its Online Learning", Neurocomputing, vol. 262, pp.41-56, 2017.
[27] S. Babichev, V. Lytvynenko, A. Gozhyi, M. Korobchinskyi, "Fuzzy model of gene expression profiles reducing based on the complex use of statistical criteria and Shannon entropy", Advances in Intelligent Systems and Computing, vol. 754, pp. 567–576, 2018.
[28] J. Sabates-Bellver, L.G. Van der Flier, M. de Palo, E. Cattaneo, et al., "Transcriptome profile of human colorectal adenomas", Mol. Cancer. Res., vol. 5(12), pp. 1263–1275, 2007.
[29] T.A. Wallace, R.L. Prueitt, M. Yi, T.M. Howe, et al., "Tumor immunobiological differences in prostate cancer between African-American and European-American men", Cancer Res., vol. 68(3), pp. 927–936, 2008.
[30] A. Sanchez-Palencia, M. Gomez-Morales, J.A. Gomez-Capilla, V. Pedraza, et al., "Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer", Int. J. Cancer, vol. 129(2), pp. 355–364, 2011.

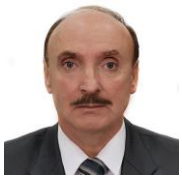
Authors' Profiles

Sergii Babichev graduated (M.Sc.) from Kherson State Pedagogical Institute in 1984. He got his PhD in 2003.

He is currently working as Associate Professor of Department of Informatics at Jan Evangelista Purkyně University in Ustí nad Labem, Czech Republic. He has about 100 scientific publications. His research interests are data mining of complex data, bioinformatics, gene expression profiles processing, gene regulatory network reconstruction and simulation.

Maksym Korobchynskyi is Doctor of Technical Sciences since 2014. From 2016, he is a Professor of Military-Diplomatic Academy named after Eugene Bereznyak, Kyiv, Ukraine. He has about 160 scientific publications. His research interests are information technology, gene expression profiles processing, management of complex dynamic objects, gene regulatory network reconstruction and robotics.

Serhii Mieshkov is Candidate of Technical Sciences (PhD) since 2014. From 2016, he is a First vice-rector of Military-Diplomatic Academy named after Eugene Bereznyak, Kyiv, Ukraine. He has about 56 scientific publications. His research interests are information technology, management of complex dynamic objects, data mining of complex data.



Oleksandr Korchomnyi is Candidate of Technical Sciences (PhD) since 2008. From 2010, he is an Associate Professor of Military-Diplomatic Academy named after Eugene Bereznyak, Kyiv, Ukraine. He has about 38 scientific publications. His research interests are data mining of complex data, management of complex dynamic objects.

How to cite this paper: Sergii Babichev, Maksym Korobchynskyi, Serhii Mieshkov, Oleksandr Korchomnyi, "An Effectiveness Evaluation of Information Technology of Gene Expression Profiles Processing for Gene Networks Reconstruction", International Journal of Intelligent Systems and Applications(IJISA), Vol.10, No.7, pp.1-10, 2018. DOI: 10.5815/ijisa.2018.07.01

