Techniques of DNA Microarray Data Pre-processing Based on the Complex Use of Bioconductor Tools and Shannon Entropy

Sergii Babichev[1,2][0000-0001-6797-1467], Bohdan Durnyak[2][0000-0003-1526-9005], Valeriy Zhydetskyy[2][0000-0002-2880-9616], Iryna Pikh[2][0000-0002-9909-8444], Vsevolod Senkivskyy[2][0000-0002-4510-540X]

[1]Jan Evangelista Purkyne University in Usti nad Labem, Ceske mladeze, 8, Usti nad Labem, 40096, Czech Republic
sergii.babichev@ujep.cz

[2]Ukrainian Academy of Printing, Pid Holoskom, 19, Lviv, 79000, Ukraine
durnyak@uad.lviv.ua, v_uad@ukr.net, pikhirena@gmail.com, senk.vm@gmail.com

Abstract. The paper presents the comparison analysis of various techniques of DNA microarrays data pre-processing in order to choose the optimal combination of the methods in terms of the minimum value of Shannon entropy criterium. The Bioconductor package tools of R software were used during the simulation process. The DNA microarray data of patients, which were investigated on lung cancer from database Array Express, were used as the experimental data. The algorithm of step-by-step procedure of the data processing for purpose of determination of the optimal combination of the methods has been proposed as the results of the research. The results of the simulation have shown that the optimal combination of the methods for the investigated data is the following one: *rma* method background correction, *invariant set* method normalization and *mas* method PM correction and summarization. This combination of the methods corresponds to the minimum value of the Shannon entropy criterion.

Keywords: DNA microarray, gene expression profiles, background correction, normalization, PM correction, summarization, Shannon entropy

1 Introduction

One of the current directions of modern bioinformatics is reconstruction and simulation of gene regulatory networks based on the data from DNA-microchip experiments [1]. Implementation of this process involves the preliminary experimental data pre-processing in order to form the array of gene expression profiles. DNA microchip data are presented as a matrix of light intensities, the values of which are proportional to the expression of the appropriate genes. The genes expression value determines the amount of appropriate type of protein which will be generated by this gene. Each of the DNA microchips includes the results of the

experiment for appropriate investigated object or for appropriate conditions of the experiment performing. In this case the quantity of the DNA microchips corresponds to the quantity of the investigated objects. It should be noted, that conditions of the experiments performing are differed in the most cases. In this case very important is the stage of the obtained data pre-processing. This procedure involves the use of four steps: background correction, normalization, PM correction and summarization [2–5]. Each of the steps can be implemented using various methods.

The technique to determine the optimal combination of methods to form the gene expression profiles array objectively is absent nowadays. In this paper we propose the technique of gene expression array formation based on Shannon entropy criterion which is calculated with the use of James-Stein shrinkage estimator method [6]. The optimal combination of the methods is determined based on the minimum value of the Shannon entropy criterion.

2 Formal problem statement

The data, which are obtained during the DNA microchip experiments, are presented as a matrix of light intensities. A block chart of procedure of DNA microchip light intensities matrix formation during the experiment performing is presented in fig. 1.

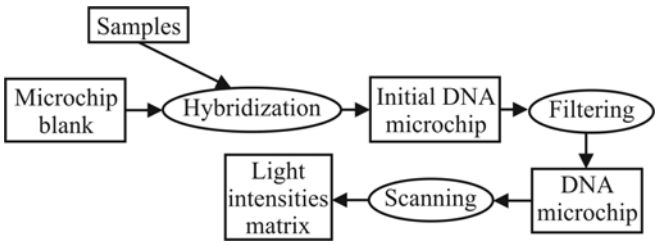

Fig. 1. A block chart of the procedure of DNA microchip light intensities matrix formation

Joining of complementary single-chain nucleotides with fluorescent labels to a single molecule is performed during the hybridization process. It is obvious, that the level of light intensities in appropriate point of the microchip is proportional to quantity of the hybridized RNA molecules, which correspond to appropriate type of the protein. The following stages of the DNA microchip processing are filtering in order to remove unhybridized samples and scanning for purpose of the matrix of light intensities formation. Fig. 2 presents the step-by-step procedure to transform the light intensities values to the expression of the corresponding genes. As it can be seen from fig. 2, each of the steps assumes the use of various methods and choice of the combination of these methods influences directly to the quality of the obtained genes expression data. Thus, the main problem consists in determination of the optimal combination of methods to process the DNA microchip data in order to increase the informativity level of the obtained gene expression data.

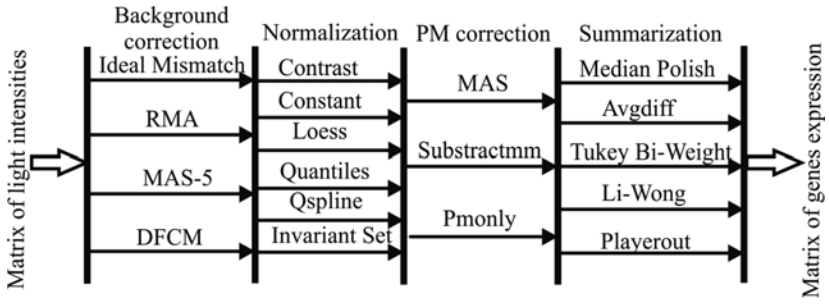

Fig. 2. A chart of step-by-step procedure to transform the light intensities matrix to the matrix of genes expression

3 Literature review

The issues concerning the DNA microarray processing are presented in [7–9]. The authors considered in detail the stages of DNA microarrays creation and the peculiarities of their processing. In [10] the author considered the possibility of the neuro-fuzzy modeling implementation for purpose of the microarrays experiments data processing. In [11] the authors presented the results of research concerning analysis of genes expression for the purpose of the objects classification using Bayesian network. However, it should be noted, that the hereinbefore works do not contain the investigations concerning choice of the appropriate combination of the methods to process the DNA microchip data based on quantitative criteria.

Classification and detail description of the background correction methods are presented in [2–5]. Ideal Mismatch method was proposed by Affymetrix company [3]. This method involves the complex use of both the Perfect Match (PM) nucleotide samples which fully correspond to the investigated genes and the Miss Match (MM) samples, in which the mean nucleotide is changed to complementary one. Robust Multichip Average (RMA) background correction method involves the use only PM samples [4]. This fact decreases the costs to the microchip preparing due to absence of the MM samples. The values of light intensities in this case are presented as the sum of the useful signal, which is distributed exponentially, and the normally distributed noise component. Distribution Free Convolution Model (DFCM) background correction method [5] also assumes that values of light intensities are presented as the combination of both the useful signal and the noise component. But in this case do not any assumes about the character of the components distribution. This method involves the use of both the PM and MM samples. The main idea and the detail description of the Affymetrix Micro Array Suite 5.0 (MAS 5.0) technique of background correction are presented in [2,3].

The techniques of DNA microchip data normalization are presented in [2,12–16]. The necessity of this stage is determined by low correlation of the data which were determined when different conditions of the experiment performing. The aim of the normalization process is the reduction of the microchip empirical data to the same distribution. This step allows minimizing the technological differences between the

parameters of different genes and, as the result, to carry out the comparison of the expression values of the corresponding genes obtained under different conditions of the experiment performing. The results of the research concerning comparison analysis of various methods of PM corrections and summarization of the DNA microchip data are presented in [2,14,17,18]. PM correction stage is performed in order to reduce the nonspecific hybridization effect by correction of the PM samples light intensities taking into account the light intensities of the corresponding MM samples. The summarization process assumes the calculation of gene expressions values from light intensities of the samples for investigated genes.

However, it should be noted that in spite to the achievements in this subject area the effective technique to choose the optimal combination of methods of DNA microchip data processing is absent nowadays. This problem can be solved based on modern techniques of the complex data processing which are used in different field of scientific research nowadays [19–23].

The aim of the paper is the improvement of technique of DNA microarray data processing based on the complex use of Bioconductor tools and Shannon entropy for purpose of gene expression array formation.

4 Materials and methods

The Shannon entropy criterion, which is calculated based on James-Stein shrinkage estimator [6], was used as the main criterion to estimate the gene expression informativity during the simulation process. This technique is based on the complex use of the two different models: a high-dimensional model with low bias and high variance, and a lower dimensional model with larger bias but smaller variance. Evaluation of the probability of values distribution in cells in accordance with James-Stein shrinkage technique is calculated by the formula:

$$p_i^{Srink} = \lambda p_i + (1-\lambda) p_i^{ML} \tag{1}$$

where p_i^{ML} is the probability of the gene expression values distribution in the *i*-th cell which is calculated by maximum likelihood method, $p_i = 1/n_i$ is the shrinkage target or probability in the *i*-th cell in the case of uniform distribution of the gene expression values, n_i is quantity of the features in the *i*-th cell. It is obvious, that p_i^{ML} corresponds to the high-dimensional model with low bias and high variance and p_i corresponds to the models with higher bias and lower variance of the features distribution. Parameter of the intensity λ in this case is calculated as follows:

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

where n is the quantity of the features in the investigated vector. The value of the Shannon entropy is calculated with the use of standard formula taking into account the method of the probability estimation:

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

Less value of the criterion (3) corresponds to the higher level of the investigated vector informativity.

A structural block chart of the algorithm which was used to determine the optimal combination of the methods of DNA microarray data processing is shown in fig. 3. Implementation of this algorithm involves the following steps:

1. Loading of the DNA microarray data.
2. Setup of the stage of data processing (background correction, normalization, PM correction, summarization). Fixation of the methods, which do not correspond to this stage randomly.
3. Choice of the first method for current stage.
4. DNA microarray data processing by selected methods.
5. Calculation of the Shannon entropy by formulas (1)–(3) for each of the microchips. Calculation of average value of the Shannon entropy for all DNA microarrays.
6. If the number of the method is less than the maximum quantity of the methods at this stage, then choice the next method and go to the step 4 of this procedure. Otherwise fixation of the method which correspond to the minimum value of the Shannon entropy.
7. If the number of the stage is less than maximum quantity of the stages, then go to the next stage and go to the step 3 of this procedure. Otherwise, DNA microarray data processing with the use of determined combination of the methods.

5 Experiments

Simulation process of DNA microchip data pre-processing was performed based on R software [24] using functions of *Bioconductor* package [25]. The lung cancer patients' gene expression profiles E-GEOD-68571 [26] from database ArrayExpress [27] were used as the experimental data during the simulation process. These data includes 96 of DNA microchips of patients which were investigated on lung cancer.

Each of the DNA microchips includes 7129 of genes. 10 patients were identified as healthy and 86 sick patients were divided by the state of their health into three groups.

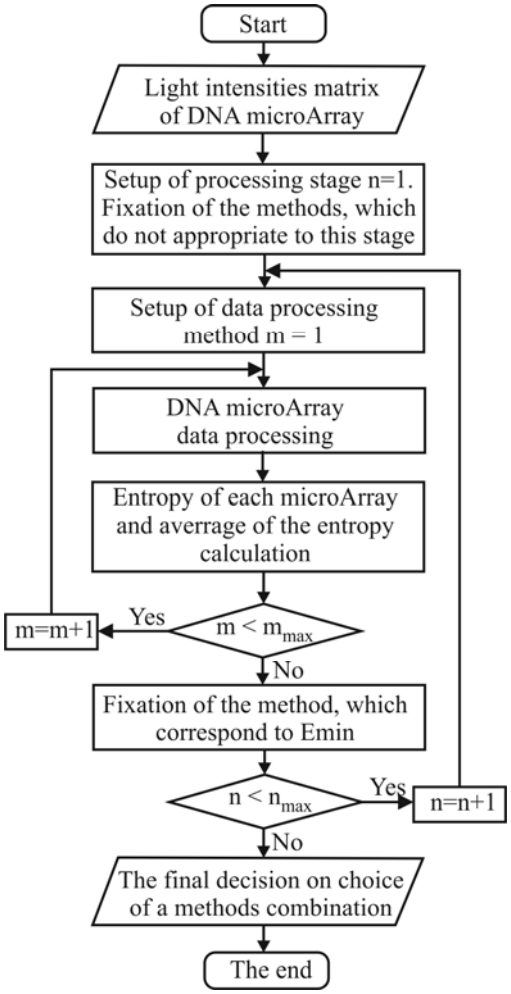

Fig. 3. A chart of the step-by-step procedure to transform the light intensities matrix to the matrix of genes expression

Fig. 4 shows the images of nine of the investigated DNA microchips which were obtained by scanning of the appropriate objects. The following step of the data processing is the transform of the light intensity matrixes into array of genes expressions using the hereinbefore presented methods.

6 Results and discussions

The character of light intensities values distribution at the selected DNA microarray is presented in fig. 5. Fig. 6 shows the MA charts for all pairs of five selected DNA microchips. The *MA* chart presents the difference of logarithms of the PM (Perfect Math) samples values (*M*) versus the average of logarithms of the PM samples values (*A*). Parameters *M* and *A* for *i*-th gene and samples *k* and *n* are calculated in the following way:

$$M = \log_2\left(\frac{x_{ki}}{x_{ni}}\right), \quad A = \frac{1.2}{}\log_2\left(x_{ki} \cdot x_{ni}\right) \tag{4}$$

The chart is created for PM values for all possible pairs of the investigated samples. In the case of the highest quality of the data processing, the data should be distributed in a rather narrow range, and the points at *MA* diagram should be located along the axis of $M = 0$ with the lowest averages.

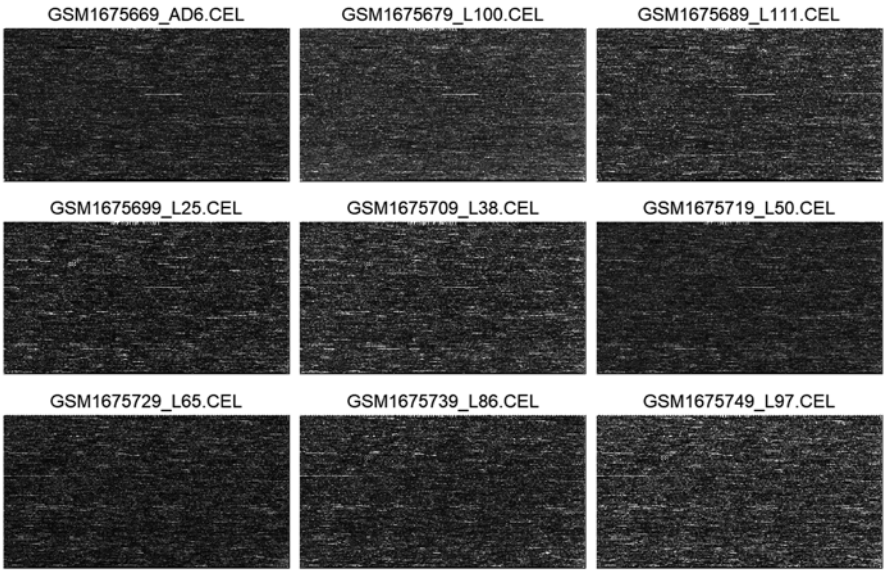

Fig. 4. Scanning images of nine of the investigated DNA microchips

The analysis of the received diagrams confirms the assumption concerning the necessity of the initial data preprocessing. The character of the data distribution for various microchips is differed significantly (Fig. 5a). The kernel density plots which are shown in fig. 5b are distributed along axis of the light intensities logarithm randomly too. Finally, the corresponding points on the MA diagrams (Fig. 6) have different distributions too. These facts do not allow us to compare the investigated gene expression profiles objectively. Fig. 7 and Fig. 8 present the results of the research concerning background correction of the DNA microarrays by methods:

"rma", "mas" and "DFCM". The "Ideal Mismatch" method has not used due to lower quality of its operation [28]. The analysis of the obtained charts allows us to conclude that the background correction increases the image quality. The processed data are distributed more uniformly to compare with unprocessed data. However, it should be noted that visual analysis of the diagrams does not allow us to compare the quality of the used methods objectively in order to choose the best one. Fig. 9 presents the results of the research concerning determination of the optimal combination of the methods to process the DNA microarray data based on the minimum value of the Shannon entropy in accordance with hereinbefore technique.

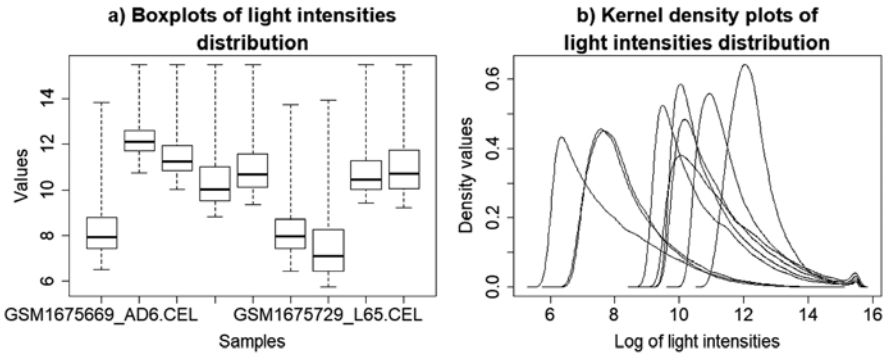

Fig. 5. Estimation of light intensities distribution at the selected DNA microchips

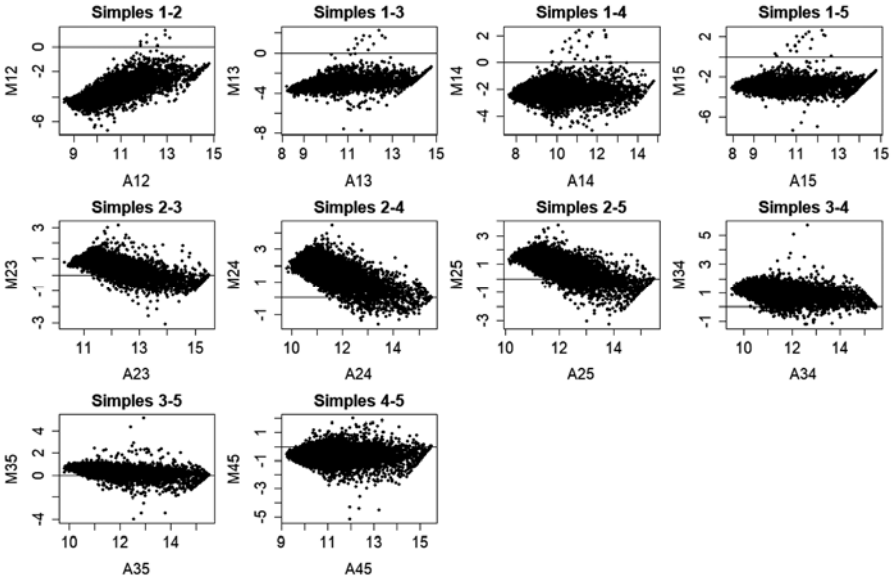

Fig. 6. *MA* charts of light intensities distribution for PM samples

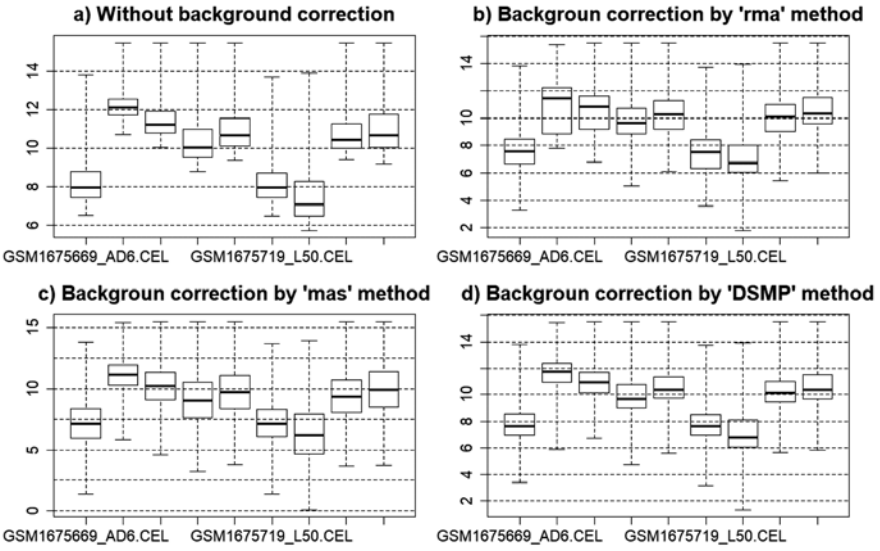

Fig. 7. Boxplot charts of unprocessed and processed data when the various background correction methods are used

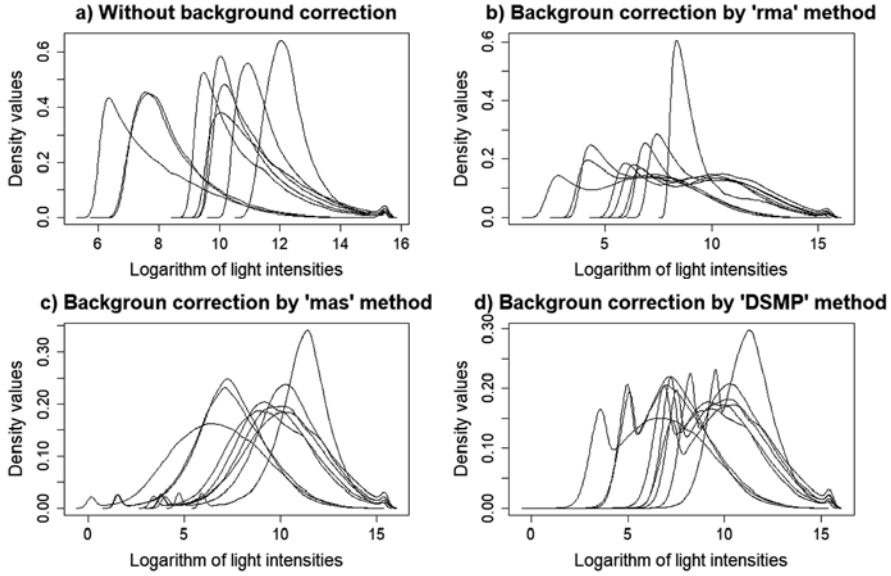

Fig. 8. Kernel density plots of unprocessed and processed data when the various background correction methods are used

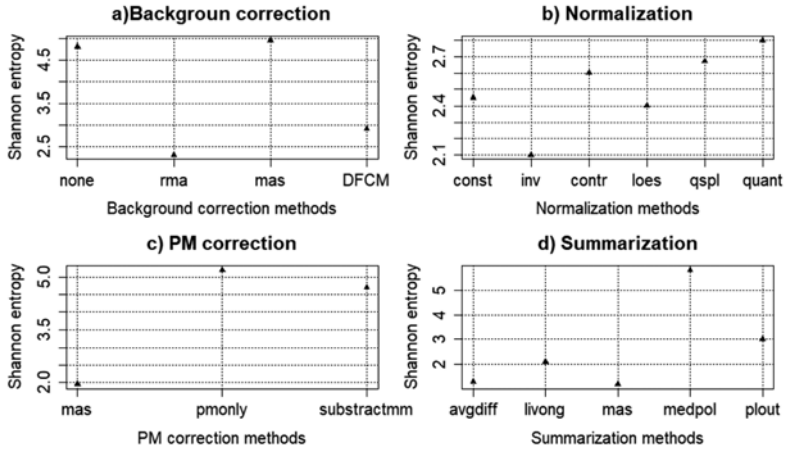

Fig. 9. Charts of the Shannon entropy distribution versus the methods of the data processing at the stages: a) background correction; b) normalization; c) PM correction; d) summarization

The analysis of the obtained charts allows us to conclude that the optimal methods in terms of the minimum value of Shannon entropy criterion are the following ones: "*rma*" background correction method; "*invariant set*" normalization method; "*mas*" methods PM correction and summarization. This combination of the methods was used to process the investigated DNA microarrays. Fig. 10 presents the boxplots of genes expression vectors for the investigated samples of both the non-processed (fig. 10a) and processed (fig. 10b) data.

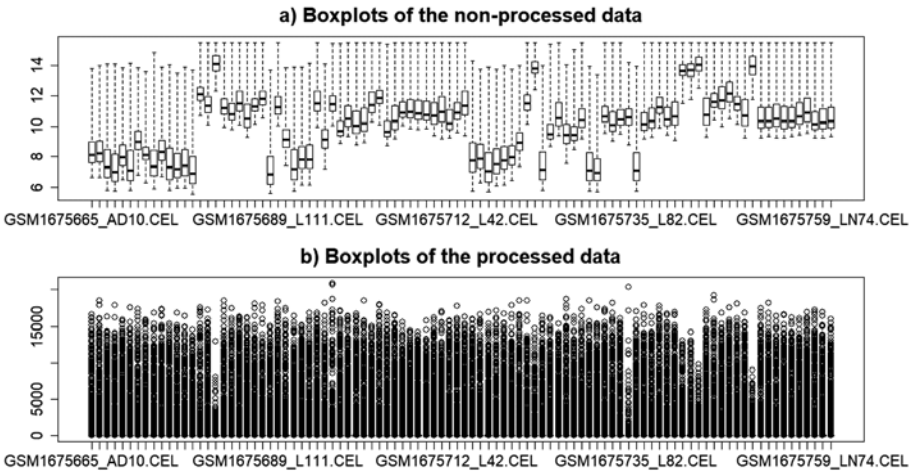

Fig. 10. Results of the DNA microchips processing

As it can be seen from fig. 10b, the values of genes expression are distributed in the same range. The change of this range can be explained in the following way. The expression values of the largest quantity of genes are low. But some of the genes have significantly higher values of expression. It means that these genes determine some important processes in the investigated objects. The expression values of these genes determine the variation range of another genes expression. The analysis of the boxplots allows us also to conclude that the values of the largest quantity of gene expressions for various objects lie in a very narrow range. This can mean that these genes are responsible for the functions that are inherent for all investigated objects. However, each of the investigated samples contains genes, the expression of which goes beyond the inter-quartile range. These genes are very important for the following research since they allow us to distinguish the investigated objects by their particularities.

7 Conclusions

In this paper we have proposed the technique of gene expression array formation which were obtained based on DNA microarray experiments. The initial data is presented as a set of DNA microchips, each of which contains the matrix of light intensities, the values of which are proportional the expression values of the appropriate genes. Four stages have been performed during the simulation process: background correction, normalization, PM correction and summarization. Each of the stage assumed the use of different methods. The Shannon entropy criterion which is calculated based on James-Stein shrinkage estimator has been used as the main criterion to estimate the genes expression informativity.

The simulation process has been performed based on R software with the use of *Bioconductor* package functions. The lung cancer patients' gene expression profiles E-GEOD-68571 from database ArrayExpress have been used as the experimental data during the simulation process. The results of the simulation have shown that the optimal combination of the methods in terms of the minimum value of the Shannon entropy is the following one: "*rma*" background correction method, "*invariant set*" normalization method and "*mas*" methods PM correction and summarization. This combination of the methods has been used to process the investigated DNA microchips.

The boxplots of both the non-processed and processed data have been created as the simulation results. The analysis of the obtained results has shown that the values of the largest quantity of gene expressions for various objects lie in a very narrow range. It means that these genes are responsible for the functions that are inherent for all investigated objects. However, each of the investigated samples contains genes, the expression of which goes beyond the inter-quartile range. This fact can mean that these genes are very important for the following research since they allow us to distinguish the investigated objects by their particularities.

References

1. Zak, D.E., Vadigepalli, R., Gonye, G.E., Doyle, F.J., Schwaber, J.S., Ogunnaike, B.A.: Unconventional systems analysis problems in molecular biology: A case study in gene regulatory network modeling. Computers and Chemical Engineering, 29 (3), pp. 547-563 (2005) doi: 10.1016/j.compchemeng.2004.08.016
2. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19 (2), pp. 185-193 (2003) doi: 10.1093/bioinformatics/19.2.185
3. Affymetrix. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA, pp. 1-27 (2002)
4. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Selected Works of Terry Speed, pp. 601-616 (2012) doi: 10.1007/978-1-4614-1347-9_15
5. Chen, Z., McGee, M., Liu, Q., Kong, M., Deng, Y., Scheuermann, R.H.: A distribution-free convolution model for background correction of oligonucleotide microarray data. BMC Genomics, 10 (SUPPL. 1), art. no. S19 (2009) doi: 10.1186/1471-2164-10-S1-S19
6. Hausser, J., Strimmer, K.: Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. Journal of Machine Learning Research, 10, pp. 1469-1484 (2009)
7. Kohane, I.S., Kho, A.T., Butte, A.J.: Microarrays for an integrative genomics. Cambridge, Massachusetts, England: A Bradford book, the MIT press, 236 p. (2003)
8. Ivakhno, S.S., Kornelyuk, A.I.: Microarrays: Technologies overview and data analysis. Ukrain'skyi Biokhimichnyi Zhurnal, 76 (2), pp. 5-19 (2004)
9. Babichev, S.A., Kornelyuk, A.I., Lytvynenko, V.I., Osypenko, V.V.: Computational analysis of microarray gene expression profiles of lung cancer. Biopolymers and Cell, 32 (1), pp. 70-79 (2016) doi: 10.7124/bc.00090F
10. Wang, Z.: Neuro-Fuzzy Modeling for Microarray Cancer Gene Expression Data. Oxford University Computing Laboratory, 107 p. (2005)
11. Loren van Themaat, E.V.: On the Use of Learning Bayesian Networks to Analyze Gene Expression Data: Classification and Gene Network Reconstruction. University of Amsterdam, 73 p. (2005)
12. Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S.: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, 473 p. (2005)
13. Park, T., Yi, S.-G., Kang, S.-H., Lee, S.Y., Lee, Y.-S., Simon, R.: Evaluation of normalization methods for microarray data. BMC Bioinformatics, 4, art. no. 33, 13 p. (2003) doi: 10.1186/1471-2105-4-33
14. Raddatz, B.B., Spitzbarth, I., Matheis, K.A., Kalkuhl, A., Deschl, U., Baumgärtner, W., Ulrich, R.: Microarray-Based Gene Expression Analysis for Veterinary Pathologists: A Review. Veterinary Pathology, 54(5), pp. 734-755 (2017) doi: 10.1177/0300985817709887
15. Åstrand, M.: Contrast normalization of oligonucleotide arrays. Journal of Computational Biology, 10 (1), pp. 95-102 (2003) doi: 10.1089/106652703763255697
16. Chen, Y.-J., Kodell, R., Sistare, F., Thompson, K.L., Morris, S., Chen, J.J.: Normalization methods for analysis of microarray gene-expression data. Journal of Biopharmaceutical Statistics, 13 (1), pp. 57-74 (2003) doi: 10.1081/BIP-120017726
17. Barbará, D., Wu, X.: An Approximate Median Polish Algorithm for Large Multidimensional Data Sets. Springer-Verlag London Ltd. Knowledge and Information Systems, vol. 5, pp. 416-438 (2003)

18. Lazaridis, E.N., Sinibaldi, D., Bloom, G., Mane, S., Jove, R.: A simple method to improve probe set estimates from oligonucleotide arrays. Mathematical Biosciences, 176 (1), pp. 53-58 (2002) doi: 10.1016/S0025-5564(01)00100-6
19. Babichev, S., Lytvynenko, V., Osypenko, V.: Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, art. no. 8098832, pp. 479-484 (2017) doi: 10.1109/STC-CSIT.2017.8098832
20. Babichev, S., Korobchynskyi, M., Lahodynskyi, O., Korchomnyi, O., Basanets, V., Borynskyi, V.: Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles. Eastern-European Journal of Enterprise Technologies, 1 (4-91), pp. 19-32 (2018) doi: 10.15587/1729-4061.2018.123634
21. Babichev, S., Krejci, J., Bicanek, J., Lytvynenko, V.: Gene expression sequences clustering based on the internal and external clustering quality criteria. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 1, art. no. 8098744, pp. 91-94 (2017) doi: 10.1109/STC-CSIT.2017.8098744
22. Tkachenko, R., Doroshenko, A., Izonin, I., Tsymbal, Y., Havrysh, B.: Imbalance data classification via neural-like structures of geometric transformations model: Local and global approaches. Advances in Intelligent Systems and Computing, 754, pp. 112-122 (2019) doi: 10.1007/978-3-319-91008-6_12
23. Peleshko, D., Ivanov, Y., Sharov, B., Izonin, I., Borzov, Y.: Design and implementation of visitors queue density analysis and registration method for retail videosurveillance purposes. Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing, DSMP 2016, art. no. 7583531, pp. 159-162 (2016) doi: 10.1109/DSMP.2016.7583531
24. Ihaka, R., Gentleman, R.: R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5 (3), pp. 299-314 (1996) doi: 10.1080/10618600.1996.10474713
25. El. Resource: https://www.bioconductor.org/about/
26. Beer, D.G., Kardia, S.L.R., Huang, C.-C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M.G., Iannettoni, M.D., Orringer, M.B., Hanash, S.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, 8 (8), pp. 816-824 (2002) doi: 10.1038/nm733
27. El. Resource: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-6857/
28. Baldi, P., Hatfield, G.W.: DNA Microarrays and gene expression: From experiments to data analysis modeling. Cambridge University Press, pp. 22-23 (2002)