ISSN 1993-6842 (on-line); ISSN 0233-7657 (print)
Biopolymers and Cell. 2016. Vol. 32. N 1. P 70–79
doi: http://dx.doi.org/10.7124/bc.00090F

UDC 651.3:678

Computational analysis of microarray gene expression profiles of lung cancer

S. A. Babichev[1], A. I. Kornelyuk[2], V. I. Lytvynenko[3], V. V. Osypenko[4]

[1] Jan Evangelista Purkyne university in Usti nad Labem, Department of Science, 40096, Cheske mladeze 8, Usti nad Labem, Czech Republic

[2] Institute of Molecular Biology and Genetics, NAS of Ukraine, 150, Zabolotnogo Str., Kyiv, Ukraine, 03680

[3] Kherson National Technical University, 24, Beryslavske highway, Kherson, Ukraine, 73008

[4] National university of life and environmental sciences of Ukraine, 15, Geroiv Oborony Str., Kyiv, Ukraine, 03041

sergii.babichev@ujep.cz, kornelyuk@imbg.org.ua, immun56@gmail.com, vvo7@ukr.net

Aim. The article is dedicated to optimization of the DNA microarray data processing, which is aimed at improving the quality of object clustering. **Methods.** Data preprocessing was performed with program R using Bioconductor package. Modelling the clustering process was made in the software environment KNIME using the program WEKA functions. **Results.** The data preprocessing is shown to be optimal while using such techniques as the background correction rma method, quantile normalization, mas PM correction and summarization by mas method. The simulation results have demonstrated a high effectiveness of the clustering algorithm Sota for this category of data. **Conclusion.** Improvement of the quality of biological object clustering is possible by means of hybridization and optimization of the methods and algorithms at different stages of data processing.

K e y w o r d s: clustering, gene expression, data preprocessing, DNA microchip.

Introduction

The DNA microarray technology is one of the modern areas of molecular biological research that allows us to identify and carry out quantitative analysis of thousands of genes simultaneously [1–3]. Qualitatively made an analysis of gene expression of the studied biological organism contributes to the determination of mechanism of disease at an early stage, and the nature of gene expression change allows to predict the nature of the appropriate type of disease further development.

There are many international computer databases of biological research objects of various nature (Array Express *etc.*). Singularity of the DNA microarray data is a large dimension of the feature space (~ 100000), that describes the object and a high level and diversity of the noise component. The emergence of the noise component is conditioned by the influence of technological factors on the process of microchip manufacturing, reading information from microarray and its subsequent processing. The efficient processing of DNA microarray data is possible due to the improvement and the development of new methods (background correction, normalization, filtration, summarization), an optimal technology of the feature space dimension reduction and the development of new clustering and classification technol-

ogies of biological objects based on an integrated use of modern technologies and methods of system analysis and data mining.

The issues of DNA microarrays processing are presented in [4–6]. The authors consider in detail the stages of DNA microarrays creation and the peculiarities of their processing. In [7] the possibility of using the neuro-fuzzy modeling to process the results of microarray experiments has been considered for the purpose of cancer diagnostics. In [8] for the analysis of the gene expression level to create the objects classification Bayesian network was used, that allows taking into account the probabilistic character of the objects distribution in multidimensional space. Notably, despite some progress in this area, it is still actual to achieve the desired precision of classification or to get the unambiguous interpretation of the results of DNA microarray obtained by clustering of investigated objects.

The aim of the paper is to research into the ways of optimizing methods of DNA microarray data processing, which is aimed at improving the quality of object clustering.

Materials and Methods

The structural flowchart of the processing of the light intensities matrix of corresponding genes of investigated objects is shown in Fig. 1. The need of background correction is caused by imperfection of the received image scan system. In this paper two methods of background correction were used: Affymetrix MicroArray Suite (MAS) method of MisMatch(MM) and Perfect Match(PM) tests proposed by company Affymetrix [9] and Robust Multi-Analysis (RMA) method [10]. In the first case MM and PM samples are used, herewith each chip is broken into 16 parts.

For each part used the least 2 % of intensity for the background correction of respective areas. The second method assumes that the light intensity at appropriate point consist of useful component and noise that have normal distribution. If we assume that α – average value exponentially-distributed signal, μ and σ^2 – the mathematical expectation and variance of the noise

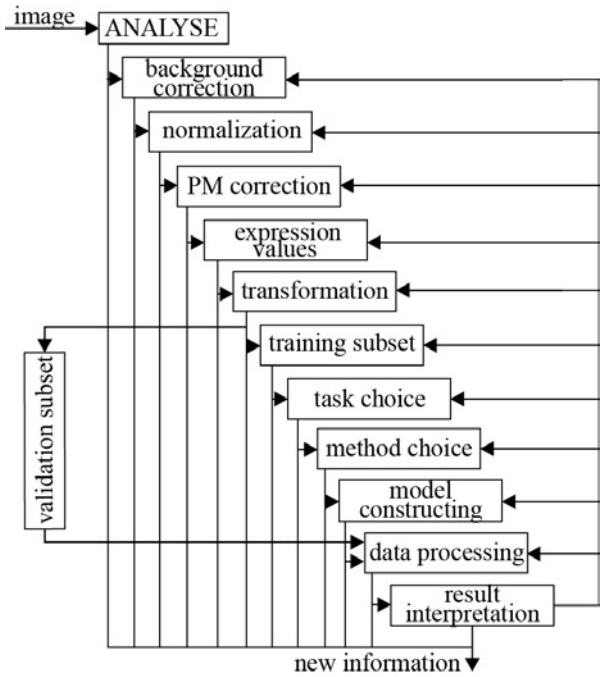


Fig. 1. Structural flowchart of an image processing of DNA microchip

component respectively, the signal intensity correction in the corresponding point occurs with the formula:

$$S = a + b\frac{\varphi\left(\frac{a}{b}\right) - \varphi\left(\frac{I-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(\frac{I-a}{b}\right) - 1} \quad (1)$$

where $a = I - \mu - \sigma^2\alpha$; $b = \sigma$; Φ and φ – distribution function and the density of the standard normal distribution respectively.

The process of data normalization leads to their single range that allows to carry out a comparative analysis of the research objects for the purpose of their classification or clustering in future. In this paper the following normalization methods were used:

- constant or scaling normalization, proposed by company Affymetrix [9]. When using this method all the arrays are scaled so that they have the same average value of light intensities;
- loess normalization [11]. Method provides calculation of variables for all values of light intensities of each pair of microarray:

$$M_i = \log_2\left(\frac{x_{ik}}{x_{in}}\right) \quad (2)$$

$$A_i = \log_2(x_{ik} \cdot x_{in}) \quad (3)$$

where x_{ik} and x_{in} – intensity value of i-th test on k-th and n-th microarrays respectively. It is assumed that between vectors A and M exists the regressive dependence. Further the normalizing correction is calculated:

$$\delta M_i = M_i - \hat{M}_i \quad (4)$$

where $\hat{M}_i$ – values of the regression function that corresponds to i-th sample.

Normalizing values of intensities are calculated as follows:

$$x'_{ik} = 2^{\left(A_i + \frac{\delta M_i}{2}\right)}, \quad (5)$$

$$x'_{in} = 2^{\left(A_i - \frac{\delta M_i}{2}\right)} \quad (6)$$

- contrast data normalization [12]. Calculation of vectors A and M and use of regression dependence between them are going to be investigated, but not all pairs of arrays are analyzed, on the first step the basic array is chosen and the whole set of calculations is performed in accordance to it;
- invariant set normalization [13]. The method assumes the use of a basic subset of PM samples as possible low light intensity distribution within each sample. Next a nonlinear relationship between light intensity values in the basic subset of samples and in investigated samples is found. This dependence later [is] used to normalize the data;
- qspline data normalization [14]. Used cubic splines and quantiles of light intensities of corresponding arrays. Spline interpolation between the respective quantiles of investigated data is realized in the normalization process;
- quantiles data normalization [15]. To use this method the projection of all points of n-dimensional quantile space on diagonal that defined by the unit vector $\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$ is calculated. In the case of direct diagonal, the line of light intensities values in all microarrays will be distributed equally.

PM correction is performed to reduce the effect of nonspecific hybridization that contributes to the noise component of investigated data. In this paper we used "mas" and "substractum" PM correction that was offered by company Affymetrix [9] and "pmonly" PM correction.

During summarization performed the calculations of expression of corresponding gene depending on the light intensity at this point. In this paper following summarization methods have been used:

- average difference method, which allows the calculation of the level of gene expression as an average of the light intensities of corresponding samples. The method proposed by company Affymetrix [9];
- liwong method [13] based on an assumption that between the intensity values in the array PM i-th sample and j-th one, or between difference of light intensities PM-MM probes and level of expression of the corresponding gene in the array and θ_i there is the following relationship:

$$p_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + e_{ij} \quad (7)$$

where $\sum \phi_i^2 = J$ – the number of simple pairs in the investigated set, e_{ij} – random error. Estimation of expression of the corresponding gene was calculated as the weighted average value of the difference of PM-MM:

$$\hat{\theta}_i = \frac{\sum p_{ij}\,\phi_j}{J} \quad (8)$$

- mas method was offered by the company Affymetrix. The value expression is calculated as robust average using 1-step Tukey biweight on $\log_2$ scale;
- median polish method [10] takes into account the difference in the nature of the interaction of genes with samples. Method based on the following additive models:

$$\log_2(y_{ij}) = \alpha_i + \mu_j + e_{ij} \quad (9)$$

where α_i – the coefficient of interactions of i-th sample with genes; μ_j – concentration of the j-th gene, which is taken as the expression of the corresponding gene.

To assess the quality of information processing Shannon entropy criterion has been used:

$$E = -\sum \theta_i \log_2 \theta_i \quad (10)$$

where θ – empirical parameter that was calculated depending on the method used.

In this paper the following methods for calculating θ has been used:

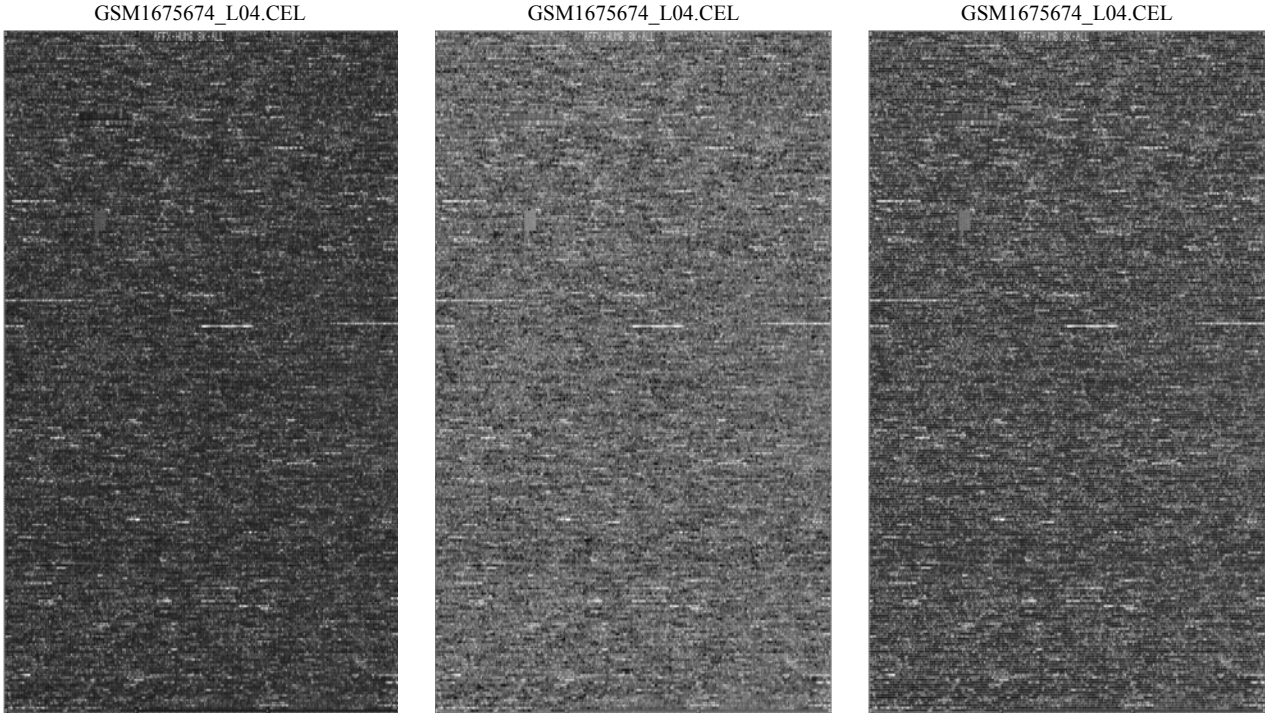
- maximum likelihood (ML);
- bias-corrected maximum likelihood (MM);
- method Dirichlet with a=1/2 (Jeffreys);
- method Dirichlet with a=1(Laplace);
- method Dirichlet with a=1/length(y) (SG);
- method Dirichlet with $a = sqrt(sum(y))/length(y)$ (minimax);
- Chaosen method (CS).

The aim of the data transformation is to reduce the dimensionality of feature space, but this process is accompanied by loss of a certain amount of useful information that affects the quality of the problem solution. In this paper to reduce the dimension of the informative features space the component analysis has been used with selection of all major important components.

The clustering of objects was made: by methods K-means and C-means using Euclidean distance as a measure of proximity between the objects and the related cluster; by neural network algorithms SOM (Self-organizing Map); by SOTA (Self-Organizing Tree Algorithm). As the measure of proximity between objects and corresponding clusters when using SOTA algorithm, Euclidean and cosine distances have been used.

Results and Discussion

Simulation of DNA microarrays data processing system was implemented by system KNIME using programmatic functions of environment WEKA and package Bioconductor of program R.



a) none *b*) mas method *c*) rma method

Fig. 2. The spatial images of light intensity distribution of DNA microarray: *a*) unprocessed microarray; *b*) microarray processed by mas background correction; *c*) microarray processed by rma background correction

As the experimental base for research we used a database of patients with lung cancer E-GEOD-68571 database Array Express [16], which includes the gene-expression profiles of 95 patients, ten of which are healthy (Norm), and 85 patients divided by the degree of the disease into three groups: 23 patients with good state (Well), 41 patients with moderate state (Moderate-Md), and 21 patients with poor state (Poor). Fig. 2*a* presents the initial image of one of the objects. Fig. 2*b* and 2*c* show the results of background correction that was made by mas and rma methods respectively.

Table 1 presents the values of Shannon entropy when using different methods for calculating the entropy for unprocessed and processed images using rma and mas methods of background correction.

In this case the comparison was made by quantiles normalization of all data with further PM mas correction and liwong summarization method. Analysis of the results shown in Table 1, allows the conclusion that mas background correction method in this case is inefficient. The images processed by rma background correction method have the smallest entropy. This fact indicates a high quality in terms of useful information availability. The same conclusion may be made by analyzing the image in Fig. 2. The scatter diagram of the light intensity magnitude of the objects without data preprocessing is shown in Fig. 3. Median values of various vectors of investigated objects indicate the necessity of data normalization.

The entropy of normalized data with different methods of normalization is presented in Table 2. The data analysis in Table 2 allows a conclusion that by the Shannon's entropy criterion the usage of quantum data normalization is optimal.

Fig. 4 shows the scatter diagram of normalized data using quantiles method of data normalization. The analysis results (Fig. 4) indicate a high quality of data normalization because they have the same median and are distributed in the same range. Tables 3 and 4 show the values of Shannon's entropy when using different methods of PM correction and summarization respectively.

Data analysis of Tables 3 and 4 allows a conclusion that the data have minimum entropy when using the mas method RM correction and summarization. Thus, the optimal preprocessing stages of DNA microarray data are: rma background correction meth-

Table 1. **Shannon entropy with different methods of background correction**

Bgcorrect method	Shannon entropy						
	ML	MM	Jeffreys	Laplace	SG	Minimax	CS
none	8.1478	8.1482	8.1484	8.1489	8.1478	8.1483	8.1478
mas	8.1591	8.1595	8.1597	8.1602	8.1591	8.1596	8.1591
rma	6.4146	6.4158	6.4211	6.4274	6.4146	6.4178	6.4155

Table 2. **Shannon's entropy at various normalization methods**

Normalization method	Shannon entropy						
	ML	MM	Jeffreys	Laplace	SG	Minimax	CS
constant	6.8396	6.8417	6.8510	6.8619	6.8396	6.8438	6.8581
contrasts	7.1075	7.1083	7.1112	7.1149	7.1075	7.1097	7.1091
invariantset	7.3183	7.3186	7.3196	7.3208	7.3183	7.3196	7.3184
loess	7.0750	7.0758	7.0790	7.0828	7.0750	7.0773	7.0767
qspline	7.0739	7.0749	7.0783	7.0827	7.0739	7.0763	7.0759
quantiles	6.4146	6.4158	6.4211	6.4274	6.4146	6.4178	6.4155
quantiles robust	6.4843	6.4854	6.4902	6.4960	6.4843	6.4874	6.4851

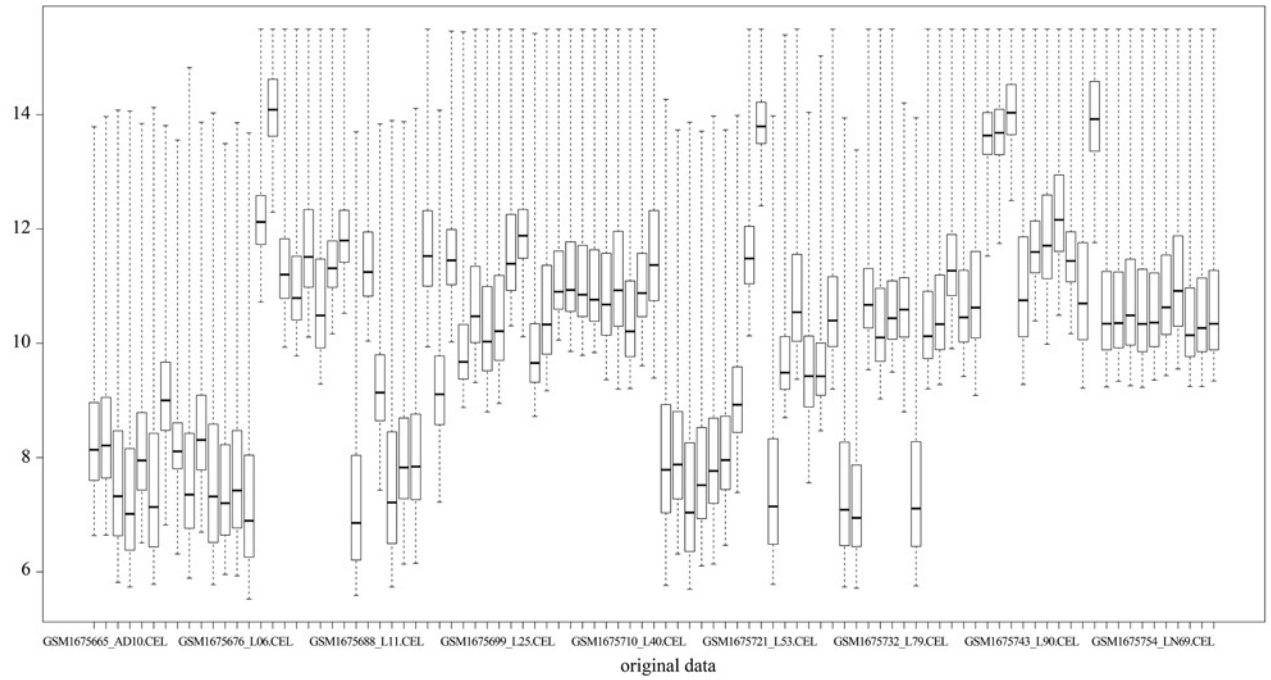


Fig. 3. Scatter diagram of not normalized data

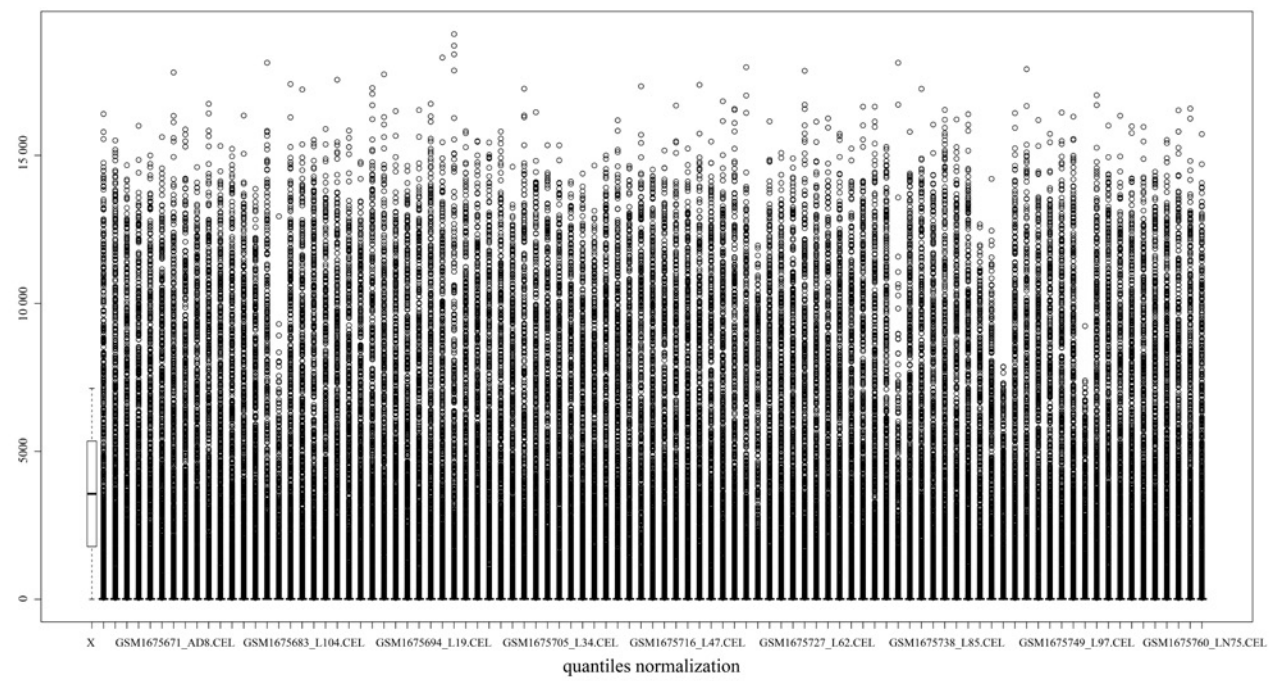


Fig. 4. Scatter diagram of normalized data

Table 3. **Shannon's entropy at various PM correction methods**

PM correction method	Shannon entropy						
	ML	MM	Jeffreys	Laplace	SG	Minimax	CS
mas	6.4146	6.4158	6.4211	6.4274	6.4146	6.4178	6.4155
pmonly	8.2268	8.2271	8.2271	8.2274	8.2268	8.2272	8.2268
substractum	11.1000	11.1000	11.1010	11.1019	11.1000	NA	6.0690

Table 4. **Shannon's entropy at various summarization methods**

Summarization method	Shannon[‘s] entropy						
	ML	MM	Jeffreys	Laplace	SG	Minimax	CS
average diff.	6.0335	6.0354	6.0447	6.0558	6.0335	6.0378	6.0356
liwong	6.4146	6.4158	6.4211	6.4274	6.4146	6.4178	6.4155
mas	5.9985	6.0004	6.0098	6.0210	5.9985	6.0029	6.0007
median polish	8.7083	8.8471	8.7443	8.7693	8.7083	8.7102	9.1911

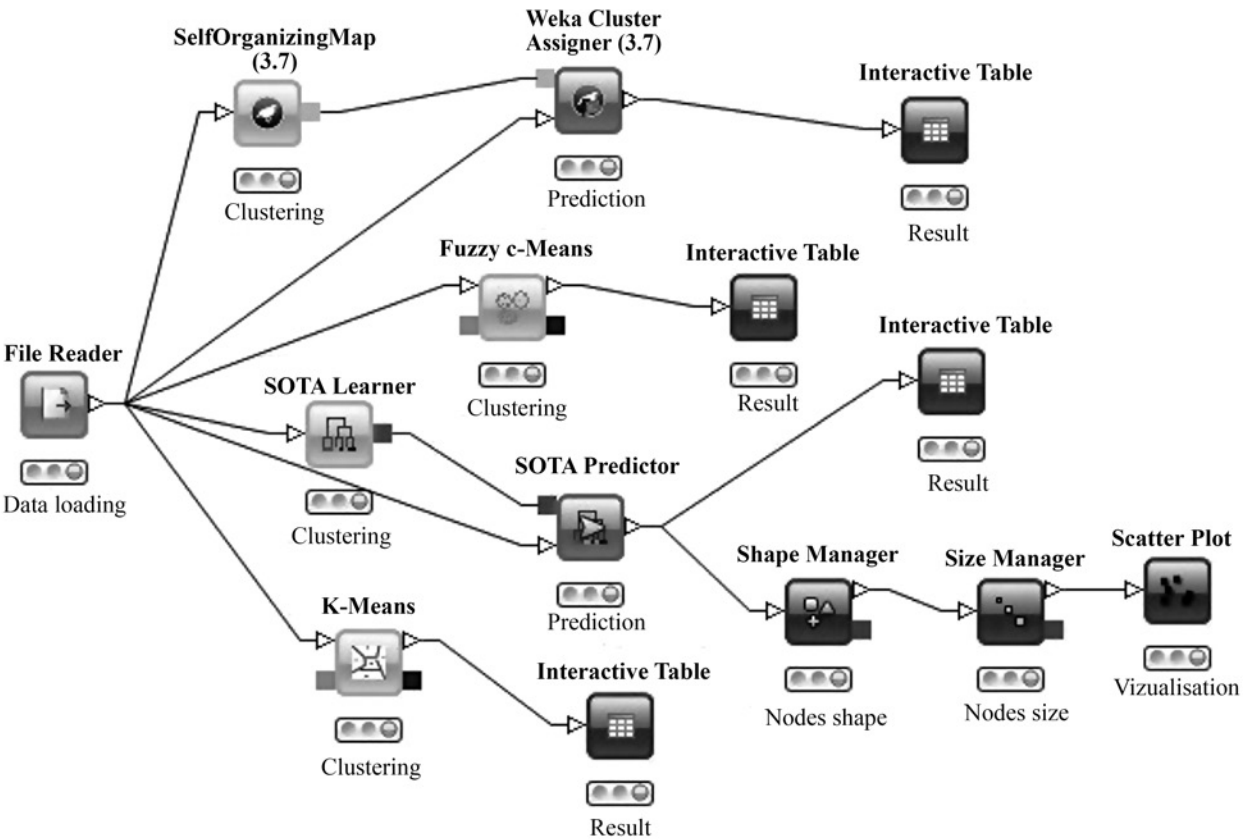


Fig.5. Model of cluster analysis of gene expression profiles

od, quantiles data normalization, mas PM correction and mas summarization method.

The data transformation was performed by calculating the principal components, and the dimension of the feature space was reduced from 7129 to 93. The model of cluster analysis for obtained database is implemented in the system KNIME and shown in Fig. 5. Table 5 presents the results of the simulation clustering process using different methods of cluster analysis.

Based on the analysis of the data in Table 5 it can be concluded that all used methods of cluster analysis share the investigated objects on patients and not patients, but in the mid of patients cluster algorithms SOM, k-means and C-means have unsatisfactory sharing resolution, because clusters Md, Well and Poor have intersection with each other. However, within the cluster of patients the algorithms SOM, k-means and C-means have unsatisfactory resolution, because clusters Md, Well and Poor intersect with each other. Additionally, some patients with good facilities condition were assign as not patients that is also unacceptable. Another conclusion can be made by analyzing the results of the algorithm Sota.

Sota algorithm clearly separates the objects on patients and not patients. Furthermore, within the cluster of patients the subclusters Poor and Well do not intersect using Euclidean distance estimation and overlap of 2.3 % using the cosine distance. There are some clusters of the objects intersection with moderate condition along with good and moderate and poor. This is logical, because the moderate patient's condition is fairly conventional concept. The patient's condition may be moderately good and moderately poor, however, the cluster of objects with moderate condition cannot be defined uniquely and it should intersect with Well and Poor clusters. Fig. 6 shows the chart of objects distribution on clusters by algorithm Sota using Euclidean distance and assessment of the degree of objects remoteness. The chart analysis confirms a good sharing ability of the algorithm Sota.

Conclusion

The research on the choice of optimal methods of DNA microarray data preprocessing has been conducted for further transformation and clustering of the investigated objects. The data preprocessing was performed using program R and consisted in the following: the correction of light intensity background in the corresponding point of the microarray; the normalization of data correction; and PM summarization, due to which the gene expression was calculated for the investigated objects.

Evaluation of the quality of information processing has been conducted using the Shannon's entropy and such methods of calculation have been used: bias-corrected maximum likelihood (MM); method Dirichlet with a=1/2 (Jeffreys); method Dirichlet with a=1(Laplace); method Dirichlet with a=1/length(y) (SG); method Dirichlet with $a = sqrt(sum(y))/length(y)$ (minimax); Chaosen method (CS). The Studies have shown that the optimal methods of data preprocessing

Table 5. **Results of cluster analysis**

Method	Norm	Norm Well	Well	Md	Poor	Md Poor	Md Well	Poor Well
Sota Euclidean	10 (100%)	0	18 (78.3%)	33 (80.5%)	15 (71.4%)	12 (19.3%)	7 (10.9%)	0
Sota Cosine	10 (100%)	0	21 (91.3%)	32 (78.1%)	15 (71.4%)	10 (16.1%)	6 (9.4%)	1 (2.3%)
SOM	10 (100%)	3 (9.1%)	10 (43.5%)	11 (26.9%)	11 (52.4%)	16 (25.8%)	21 (32.8%)	13 (29.5%)
k-means	10 (100%)	8 (24.2%)	4 (17.4%)	27 (65.9%)	10 (47.7%)	19 (30.6%)	15 (23.4%)	2 (4.5%)
C-means	10 (100%)	8 (24.2%)	2 (8.7%)	24 (58.5%)	10 (47.6%)	21 (33.9%)	17 (26.5%)	3 (6.8%)

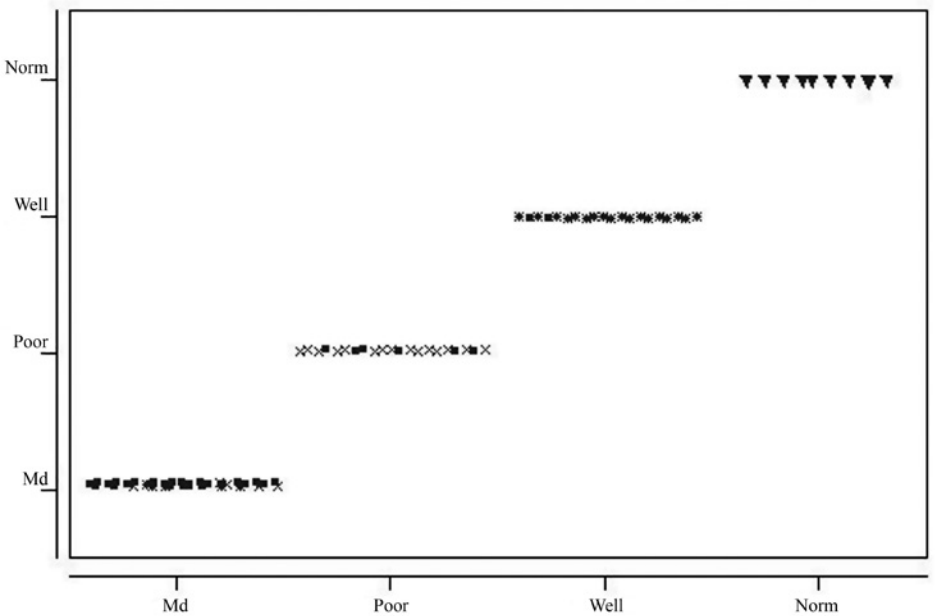


Fig. 6. Diagram of objects distribution on clusters using Sota algorithm

by the Shannon's criterion are: rma background correction method, quantiles data normalization, RM mas-correction and mas-summarization. Reducing dimension of feature space was performed by component analysis, and the primary matrix gene expressions (95×7129) were transformed to a matrix (95×93) through the selection of all significant principal components.

The modeling of cluster analysis was performed in the system KNIME using the functions of the system WEKA. The clustering algorithms Sota, SOM, k-means and C-means were investigated using the database of patients with lung cancer, 10 of whom were not patients. As a result of modeling it was established that all algorithms share the objects into the patients and not patients, but only Sota algorithm gives a satisfactory clustering result of patients within the cluster.

The prospect for further research is to develop efficient methods for clustering and classification of biological objects in order to improve the dividing ability of the presented algorithm.

REFERENCES

1. *Baldi P, Gatfield GW.* DNA microarrays and gene expression: From experiments to data analysis modeling. *Cambridge, Massachusetts, England: Cambridge University Press, 2002.* 207 p.
2. *Nepomuceno JA, Troncoso A, Nepomuceno-Chamorro IA, Aguilar-Ruiz JS.* Integrating biological knowledge based on functional annotations for biclustering of gene expression data. *Comput Methods Programs Biomed.* 2015;**119**(3):163–80.
3. *Flores JL, Inza I, Larrañaga P, Calvo B.* A new measure for gene expression biclustering based on non-parametric correlation. *Comput Methods Programs Biomed.* 2013;**112**(3):367–97.
4. *Kohane IS, Kho A, Butte AJ.* Microarrays for an integrative genomics. *Cambridge, Massachusetts, England: A Bradford book, the MIT press, 2003.* 236 p.
5. *Ivakhno SS, Korneliuk OI.* [Microarrays: technologies overview and data analysis]. *Ukr Biokhim Zh.* 2004;**76**(2):5–19.
6. *Pontes B, Giráldez R, Aguilar-Ruiz JS.* Biclustering on expression data: A review. *J Biomed Inform.* 2015. pii: S1532-0464(15)00138-0.
7. *Wang Z.* Neuro-Fuzzy modeling for microarray cancer gene expression data. *Thesis.* Oxford University Computing Laboratory, 2005. 107 p.
8. *Loren van Themaat EV.* On the use of learning bayesian networks to analyze gene expression data: classification and gene network reconstruction. *University of Amsterdam, Master Thesis 2005.* 73 p.
9. *Parrish RS, Spencer HJ 3rd.* Effect of normalization on significance testing for oligonucleotide microarrays. *J Biopharm Stat.* 2004;**14**(3):575–89.
10. *Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;**4**(2):249–64.
11. *Dudoit S, Yang YH, Callow MJ, Speed TP.* Statistical methods for identifying genes with differential expression in rep-

licated cDNA microarray experiments. *Statistica Sinica.* 2002; **12**(1): 111–28.
12. *Astrand M.* Contrast normalization of oligonucleotide arrays. *J Comput Biol.* 2003;**10**(1):95–102.
13. *Li C, Wong WH.* Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA.* 2001;**98**(1):31–6.
14. *Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 2002;**3**(9):research0048.
15. *Bolstad BM, Irizarry RA, Astrand M, Speed TP.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003;**19**(2):185–93.
16. *Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med.* 2002;**8**(8):816–24.

Комп'ютерний аналіз мікромасивів профілів експресії генів раку легенів

С. А. Бабічев, О. І. Корнелюк, В. І. Литвиненко,
В. В. Осипенко

Мета. Проведення досліджень щодо оптимізації методів, що використовуються у процесі обробки профілів експресії генів, для підвищення якості кластеризації об'єктів. **Методи.** Передобробка даних була виконана у програмному середовищі R з використанням пакету «Біокондуктор». Моделювання процесу кластеризації було зроблено у програмному середовищі KNIME з використанням функцій програми WEKA. **Результати.** Показано, що оптимальним є процес передобробки даних з використанням методів: фонова корекція rma методом, квантільна нормалізація, mas PM корекція і сумаризація mas методом. **Результати** моделювання показали високу ефективність використання для даного типу даних алгоритму кластеризації Sota. **Висновки.** Проведені дослідження показали, що підвищення якості розподілу об'єктів біологічної природи на кластери можливо за рахунок гібридизації та оптимізації використання методів і алгоритмів на різних етапах обробки даних.

Ключові слова: кластерізація, експресія генів, передобробка даних, мікрочіп ДНК.

Компьютерный анализ микромассивов профилей экспрессии генов рака легких

С. А. Бабичев, А. И. Корнелюк, В. И. Литвиненко,
В. В. Осипенко

Цель. Проведение исследований по оптимизации методов, используемых в процессе обработки профилей экспрессии генов, с целью повышения качества кластеризации объектов. **Методы.** Предобработка данных выполнялась в программной среде R с использованием пакета «Биокондуктор». Моделирование процесса кластеризации производилось в программной среде KNIME с использованием функций программы WEKA. **Результаты.** Показано, что оптимальным является процесс предобработки данных с использованием методов: фоновая коррекция rma методом, квантильная нормализация, mas PM коррекция и сумаризация mas методом. **Результаты** моделирования показали высокую эффективность использования для данного типа данных алгоритма кластеризации Sota. **Выводы.** Проведенные исследования показали, что повышение качества разделения объектов биологической природы на кластеры возможно за счет гибридизации и оптимизации использования методов и алгоритмов на различных этапах обработки данных.

Ключевые слова: кластеризация, экспрессия генов, предобработка данных, микрочип ДНК.

Received 02.11.2015