International Frontier Science Letters
ISSN: 2349-4484, Vol. 8, pp 1-8
doi:10.18052/www.scipress.com/IFSL.8.1

Submitted: 2016-03-22
Revised: 2016-05-23
Accepted: 2016-05-23
Online: 2016-06-29

Filtration of DNA Nucleotide Gene Expression Profiles in the Systems of Biological Objects Clustering

Sergii Babichev[1,a], Mohamed Ali Taif[2,b], Volodymyr Lytvynenko[2,c]

[1]Department of Informatics, Faculty of Science, Jan Evangelista Purkině University in Ústi nad Labem, České mláděže 8, 400 96 Ústi nad Labem, Czech republic

[2]Department of Informatics and Computer Science, Faculty of Cybernetics, Kherson National Technical University, Beryslavske hardway, 24, 73008, Kherson, UKRAINE

[a]sergii.babichev@ujep.cz, [b]taifmohamedali@gmail.com, [c]immun56@gmail.com

Keywords: gene expression, filtration, DNA nucleotide, thresholding, clustering

Abstract. Researches on an optimization of the filtration process of DNA nucleotides gene expression profiles are presented in the article. The data of lung cancer patients E-GEOD-68571 of Array Express database were used as experimental data. Filtration was carried out under the terms of the expression detecting of corresponding gene, herewith the variance of gene expression, the absolute value of expression and the Shannon entropy were used as criteria. The value of thresholding coefficient was estimated on the basis of average proximity measure of objects within the homogenous group and between groups. 470 columns were removed in the process of data filtering, and the matrix dimension of the test data has changed from (96×7129) to (96×6659). Estimation of the quality of information processing was performed by the comparative analysis of the clustering results of processed and unprocessed data.

Introduction

Functional genomics is one of the actual directions in the field of bioinformatics nowadays. Its main task is analyzing and implementation of transfer information mechanisms, recorded in the genome of biological cells, from gene to feature. On this basis the subsequent identification of the object state is carried out. RNA Sequencing [1] and analysis of DNA microarray data [2-5] are the basic methods of gene expression determining nowadays. Each of these methods has its advantages and disadvantages. RNA sequencing method allows us to obtain the direct information about the RNA molecule nucleotides sequence of the investigated genome that in its turn allows to determine the expression absolute value of the corresponding gene. High cost of the experiment is the main disadvantage of this technology. The low cost, the possibility of simultaneous analysis of tens of thousands genes, the technology available to practical implementation are the advantages of DNA microarray technology. High error of results obtained due to the high level and specificity of noise component that arises at the stage of microarray creating and reading information from this microarray is the main disadvantage of this technology. Thereby the development of effective methods of DNA microarray data preprocessing on the basis of modern computer methods of information processing is highly relevant. The research papers [3-6] are devoted to issues of DNA microarray data processing. The authors carry out a detailed analysis of various stages of DNA microarray creating, reading information from microarray, and post-processing in order to estimate the gene expression level of the studied objects. The article [7] presents the results of the experiments for cancer patient data clustering with the use of different clustering algorithms. The authors conducted researches to choose the optimal group of methods for data preprocessing, and the Shannon entropy was used as the main criterion in this case. However, it should be noted, that issues of data filtration in accordance with specifics of noise (non-specific hybridization) are not considerably paid attention at this works. The absence of effective filtering algorithms for DNA microarray data, focused on the removing of nonspecifically hybridized genes as well as genes that do not carry the essential information about the features of the analyzed objects are unsolved parts of the general problem.

The aim of the article is the development of step by step filtering techniques of gene expression profiles of DNA microarray data based on the complex use of different criteria to estimate the gene expression variations of biological objects.

Materials and methods

The source of the investigated data is the DNA microarray, what consists of a solid surface, on which thousands of oligonucleotide sequences are arranged and fixed in a certain order, covering all known combinations of genes mutation predisposing to a certain diseases. The process of the DNA microarray manufacturing is shown in Fig. 1. Usually, at the initial stage, two groups of samples (tested and referenced) are collected. Next, complementary DNA (cDNA) is obtained from the mRNA by means of transcriptase-polymerase chain reaction. The obtained cDNA samples are called targets. The target samples are then labeled with fluorescent dyes of different colors. Reference set of genes are labeled with green dyes (Cy3) and taste samples are labeled with red dyes (Cy5).

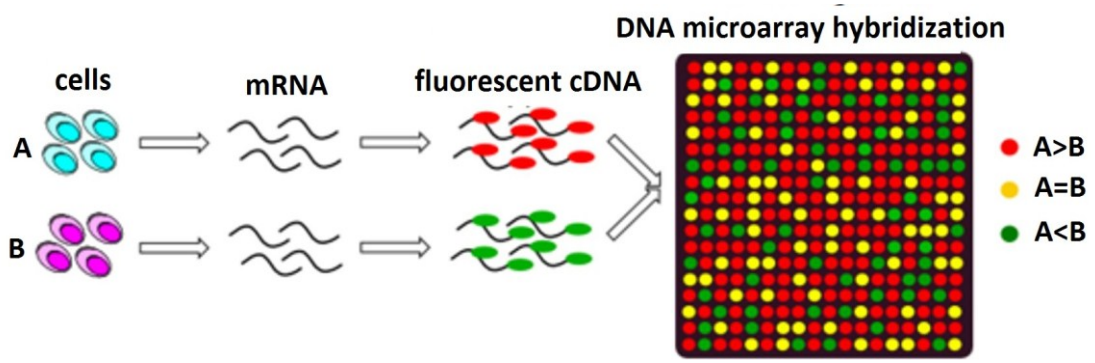

Fig. 1. Steps of DNA microarray manufacturing

Thereafter, the labeled target samples are hybridized onto the microarray according to the complementarity rule. Once the cDNA targets have been hybridized to the microarray, the array then is washed to remove any loose targets. Finally, the array is scanned by a two color laser to determine the amount of the target that is bound to each spot. Quantitatively the intensity of the light at the corresponding point is determined by the ratio of quantity of hybridized molecules of the reference and the taste substance:

$$I = \log_2 \frac{I(Cy5)}{I(Cy3)} \tag{1}$$

Thus, the matrix of light intensity in size of $(n \times m)$ is the result of DNA microarray data scanning: $A = \{x_{ij}\}, i = 1,...,n, j = 1,...,m$, where i – is the number of experiments carried out or the number of objects investigated, j – is the number of conditions under which the appropriate genes have been expressed. It is obvious that the level of gene expression should be significantly varied under the same conditions for various radically different biological objects due to the diversity of the occurring biological processes. Therefore, the genes whose expression does not correspond to this condition may be deleted from the data array as uninformative. This fact will be able to increase the resolution for further information analyses.

Filtration of DNA microarray data suggests the presence of the following stages:

– removing of genes with missing values of gene expression, arising due to the fact of some samples appeared to be unhybridized;

– removing of genes with low value of profiles variance. Low value of variance indicates insignificant change of the gene expression level during the transition from one object to another, that does not contribute to high quality of subsequent information processing;

– removing of columns with low absolute value of gene expression. Poor hybridization is the reason of the low absolute values of gene expression profiles;

– removing of genes with high absolute value of Shannon entropy [8]:

$$E_j = -\sum_{i=1}^{n} p(x_{ij}) \cdot \log_2 p(x_{ij}) \tag{2}$$

where E_j – is the entropy of j-th gene, $p(x_{ij})$ – is the probability of state realization of j-th gene of i-th object.

In accordance with the classical definition of entropy, it is a quantitative measure of the randomness of the structural elements in the system. The low value of the entropy corresponds to a high informativity due to high ordering of the gene expression level for the set of studied objects. The high entropy corresponds to a high level of disordering of expression profiles distribution for different objects that can be interpreted as noise. The threshold value of the variation feature, which defines the board of genes set division into information and non-information is determined by the condition:

$$\forall e_j: \ f(e_j) \le k \cdot \min(f(e_j)), \ k \ge 1 \ \text{ or } \forall e_j: \ f(e_j) \ge k \cdot \max(f(e_j)), \ k \le 1 \tag{3}$$

where $j = 1,...,m$ – is the number of conditions of gene expression determining, $f(e_j)$ – is the thresholding coefficient, what determines empirically in each case. The choice of thresholding coefficient value was carried out by analysis of changes of mean-square value of distance from objects to mass center of homogenous group of objects (clusters) during the removing of studied database columns:

$$D_s = \frac{1}{n_1} \sum_{i=1}^{n_1} d^2(\bar{x}_i, c_s) \tag{4}$$

and average of intercluster distance which is determined for two clusters as mean-square value of distance from objects of cluster S to mass center of cluster P and inversely:

$$D_{out} = \frac{1}{2}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} d^2(\bar{x}_i, c_s) + \frac{1}{n_2} \sum_{i=1}^{n_2} d^2(\bar{x}_i, c_p) \right) \tag{5}$$

where n_1 and n_2 – are the numbers of objects in clusters S and P respectively, c – is the mass center of corresponding cluster:

$$c = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \ , \quad j = 1,...,m \tag{6}$$

n – is the number of objects, m – is the number of attributes, that characterize objects. In the case of several clusters average intracluster distance was determined as average for intracluster distances of all clusters:

$$D_{in} = \frac{1}{q} \sum_{s=1}^{q} D_s \ , \tag{7}$$

and average intercluster distance was determined as average for intercluster distances of all pairs of clusters.

However, it should be noted that the absolute values of the criteria defined by formulas (4) - (7) have the disadvantage. The average density of the objects within the clusters distribution can be decreased during increasing of the number of removing columns (change of the thresholding coefficient k value), that will cause the increase of the average intracluster distance. Herewith, the

increase of average of intercluster distance with more high speed is possible this increase may be indicated by better quality of objects division. In this case the use of complex relative criterion, extremum of which will allow to optimize the reasonable choose of thresholding coefficient, is appropriate:

$$R = \frac{D_{out}}{D_{in}}, \tag{8}$$

As an experimental base for research we used a database of patients with lung cancer E-GEOD-68571 of the database Array Express [9], which includes the gene-expression profiles of 95 patients, ten of which are healthy (Norm). The rest 85 patients were divided by the degree of the disease into three groups: 23 patients are in good state (Well), 41 patients are in moderate state (Moderate-Md), and 21 patients are in poor state (Poor). The processing of the DNA microarray scanned image was carried out in the following way [7]: background correction by rma method, quantil normalization, mass-PM correction and summarization by mass-method. As the result, the matrix in size (96×7129) of gene expression profiles was obtained, where number of rows corresponds to the number of experiments and numbers of columns are the number of conditions of estimation gene expression profiles. There were no missing values at the obtained matrix of data. Group of objects, in which cancer cells were absent (Norm), and a group of patients with poor state (Poor) were used for estimation of filtration quality by formulas (4)-(8).

Result and discussion

Removing of genes with low values of expression variance.

The statistical values of variation of the studied genes expression variance are shown in Table 1.

Table 1.
Gene expression variance variation of investigated objects

Min	Quan_25%	Median	Mean	Quan_75%	Max
0,013	1,974	9,479	303586	111,576	29530994

The data analysis of the Table 1 allows us to conclude that variation of genes expression profiles variance of the studied objects is rather high, herewith an abundance of genes have a low expression variance, that indicates about their uninformativeness, because the values of the expression level of these genes varies insignificantly during the transition from one object to another. Genes, variance of which satisfies to the condition (9), are removed during processing:

$$\mathrm{var}(g) \le k \cdot \min(\mathrm{var}) \tag{9}$$

The value of thresholding coefficient k changed from 0 to 15 by step 0,5. Results of the experiment are shown in Fig. 2. Analysis of plots at Fig. 2a and 2b allows to conclude about inefficiency at this case of absolute values of intercluster and intracluster distances because during thresholding coefficient increase the values of these criterions monotonically increase too. However, Fig. 2c shows that value of relative distance, calculated by the formula (8) is maximum when value of thresholding coefficient is k = 14. 283 columns had been deleted in this case and new matrix of gene expression obtained dimension (96×6846). In Fig. 2d the distribution of gene expression for low and high variance is shown. In the figure it is shown that removing the column with a low gene expression variance is reasonable, because these columns are not informative for studied objects identification.

Removing of genes with low absolute values of expression level.

In the Table 2 the statistical values of average absolute genes variations are shown. The data in Table 2 allow to conclude about high variation of absolute values of studied genes. Moreover, it is stated that the most of the genes has low absolute value expression profiles, which can indicate high level of noise components.

Table 2.
Statistics of gene expression absolute values average

Min	Quan_25%	Median	Mean	Quan_75%	Max
2,182	5,032	9,8	244,278	22,571	14199

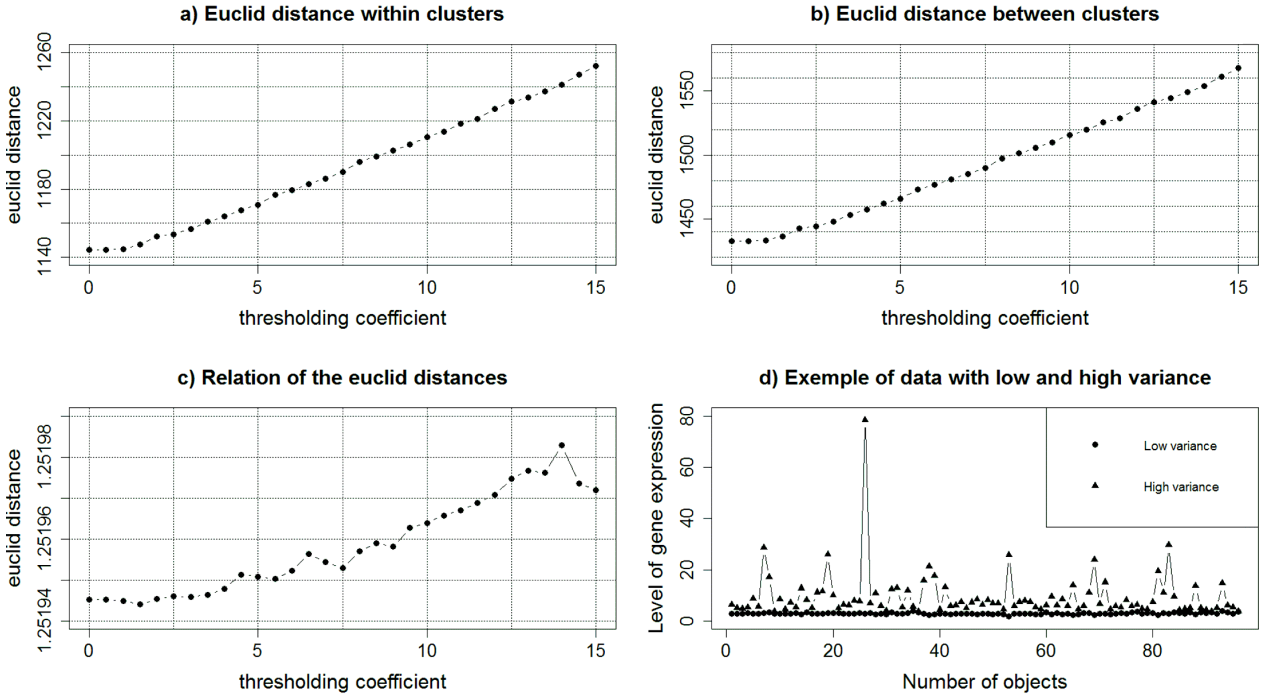

Fig. 2. Experiment results of data filtration based on using of the gene expression variance criterion: a) the plot of average intracluster distance against thresholding coefficient; b) the plot of average intercluster distance against thresholding coefficient; c) the plot of relative distance against thresholding coefficient; d) distribution of studied object genes in case of low and high variance

Removing of genes was carried out by the formula:

$$abs(g) \le k \cdot \min(abs) \qquad (10)$$

where abs – is the absolute value of the corresponding gene expression profiles. Value of thresholding coefficient k changed from 0 to 1,5 by step 0,05 during experiment. In Fig. 3 the results of experiment are shown. Analysis of plots in Fig 3 allows to conclude that the increase of thresholding coefficient value to 1,2 has no significant influence to the value of quality criteria of objects grouping. 89 columns of data were removed from matrix when k = 1,2. The further increase of thresholding coefficient value contributed to the steep increase of deleted information quantity. This fact is connected with the risk of feature space informativity decrease. Therefore the matrix of studied data takes the size of (96×6760) at this stage. Example of gene expression distribution for two objects with low and high absolute values of expression is shown in Fig. 3d. This chart also indicates low informativity of genes, expression absolute values of which for studied objects are less when installed threshold. Therefore, the delete of these columns will not significantly influence at the quality of further information processing.

Removing of genes with high values of Shannon entropy.

Table 3 shows the statistical values of Shannon entropy gene expression.

Table 3.
Statistics of entropy values of studied genes expression

Min	Quan_25%	Median	Mean	Quan_75%	Max
0,351	4,429	4,518	4,309	4,538	4,559

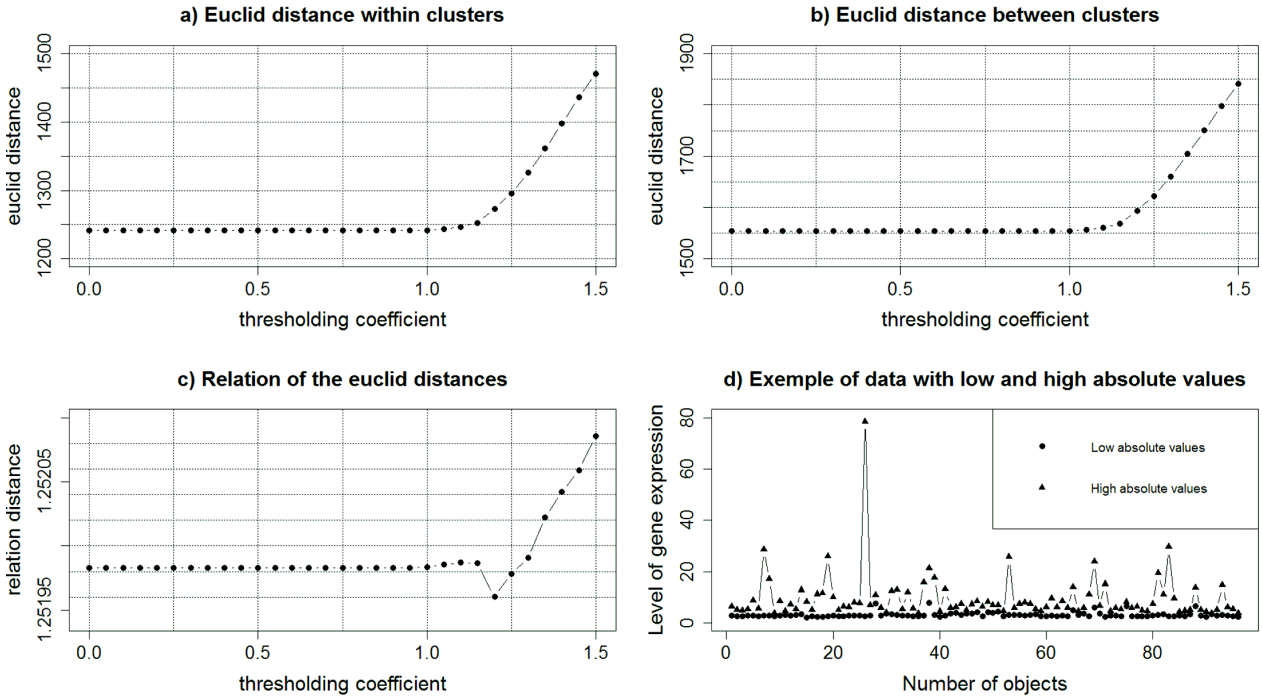

Fig. 3. Experiment results of data filtration based on using of the gene expression absolute value criterion: a) the plot of average intracluster distance against thresholding coefficient; b) the plot of average intercluster distance against thresholding coefficient; c) the plot of relative distance against thresholding coefficient; d) distribution of studied object genes in case of low and high absolute values

The analysis of the data in Table 3 shows that the majority of gene expression entropy values are displaced to the field of high values. Therefore, the value of thresholding coefficient should be insignificantly less than 1 at this case. The remove of genes was performed by the formula:

$$E(g) \geq k \cdot \max(E) \tag{11}$$

where E – is the Shannon entropy value, what have been calculated by the formula (1). The thresholding coefficient value has been changing from 0,998 to 1 by step 10^{-4} during experiment. In Fig. 4 the results of the experiment are shown. On the basis of the experiment results analysis the value of thresholding coefficient was accepted as 0,999, herewith 101 column was deleted and the matrix of new data takes the size (96×6659). Example of gene expression distribution for two objects with low and high Shannon entropy of expression is shown in Fig. 4d.

Estimation of the proposed method effectiveness was carried out by the objects clustering of initial set of normalized data, principal components of initial data set and filtering data principal components. Simulation was carried out by software KNIME using SOTA algorithm clustering [10]. All meaningful principal components were taken for component analysis and the size of array of studied objects decreased to (96×94). In Table 4 the results of simulation are shown. Data analysis of table 4 allows to draw a conclusion on effectiveness of the proposed method, because the principal components of filtering data have less clustering error at the same conditions of clustering.

The system has divided the studied objects into patient and not-patient in all cases, however the results are different within the patients group. Intersection of clusters with poor and moderate and with well and moderate states is completely logical, as for as moderate state can be both a moderate-poor and a moderate-well, but intersection of principal components of filtering data is less than in all other cases. However, clusters intersection with poor and well states is inadmissible. This intersection has not been of analyses of principal filtering data only.

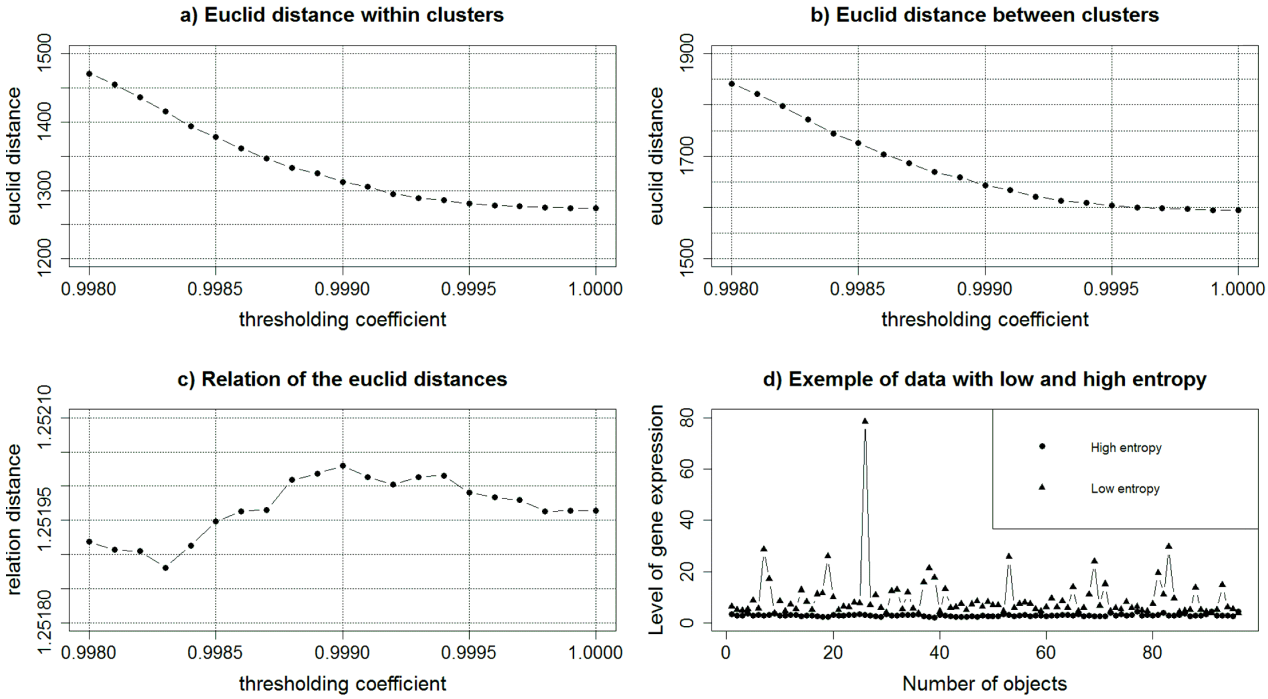

Fig. 4. Experiment results of data filtration based on using of the gene expression entropy Shannon criterion: a) the plot of average intracluster distance against thresholding coefficient; b) the plot of average intercluster distance against thresholding coefficient; c) the plot of relative distance against thresholding coefficient; d) distribution of studied object genes in case of low and high entropy

Table 4.
Results of studied data cluster analysis

Clustering error	Initial set	Denoise set	PC initial data	PC filtr. data
Well→Md	5	5	5	4
Md→Well	3	3	3	2
Poor→Md	5	5	5	4
Md→Poor	8	7	8	7
Poor→Well	1	1	1	–

Conclusion

This article presents the method of step by step DNA nucleotide filtration, obtained by DNA microarray experiments. Variance of gene expression, absolute value of expression and Shannon entropy were used as criteria to estimate the informativity of the studied objects features vectors. As an experimental base for research we used a database of patients with lung cancer E-GEOD-68571 database Array Express [9], which includes the gene-expression profiles of 95 patients, ten of which are healthy (Norm), and 85 patients are divided by the degree of the disease into three groups: 23 patients are with good state (Well), 41 patients are with moderate state (Moderate-Md), and 21 patients are with poor state (Poor). The evaluation of thresholding coefficient for removing of not-informative columns was carried out on basis of intergroup and intragroup distances calculating, herewith Euclid distance was used as a measure of proximity. 470 columns were removed during data filtration and dimension of initial array of the studied data was changed from (96×7129) to (96×6659). The cluster analysis using SOTA clustering algorithm was carried out for the evaluation of the effectiveness of the method, herewith the principal components were calculated at the preliminary stage for the purpose of feature space dimension reduction, and the number of columns of database was reduced to 94. The results of experiment have shown higher quality of filtering data principal components clustering, because the cluster intersection of objects with poor and well state inside the patients group was not observed only in this case. Moreover, the use of filtering data principal components allowed to get the best separation ability of studied objects clustering.

References

[1] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nature Reviews Genetics. 12(2011) 87-98.

[2] M. Schena, R.W. Davis, Microarray biochip technology, Eaton Publishing, 2000.

[3] P. Baldi, G.W. Hatfield, DNA Microarrays and gene expression: From experiments to data analysis modeling, Cambridge University Press, 2011.

[4] M.R. Berthold, C. Borgelt, F. Hoppner, F. Klawonn, Data Preparation, Guide to Intelligent Data Analysis, Springen-Verlag London Limited, 2010.

[5] W. Jianan, Z. Chunguang, L. Zhangxu, X. Xuefei, Z. You, L. Guixia, A Novel Workflow for Microarray Data Analysis under Expression Level of genes, Information and Computational Science. 9(2012) 4745-4754.

[6] R.A. Irizarry, B. Hobbs, F. Collin, Exploration, normalization, and summaries of high density oligonucleotide array probe level, Biostatistics. 2(2003) 249-264.

[7] S. Babichev, V. Lytvynenko, A. Kornelyuk, V. Osypenko, Computational analysis of microarray gene expression profiles of lung cancer, Biopolymers and Cell. 1(2016) 70-79.

[8] C.E. Shannon A mathematical theory of communications, Bell System Technical Journal. 27(1948) 379-423, 623-656.

[9] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nature Medicine. 8(2002) 816-824.

[10] J. Dorazo, J.M. Carazo, Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree, Journal of Molecular Evolution. 2(1997) 226-259.