UDC 004.048

THE USING OF BICLUSTERING TECHNIQUES IN INDUCTIVE MODELING SYSTEMS OF BIOLOGICAL PROCESSES

Sergii Babichev[1], Volodymyr Osypenko[2],
Taif Mohamed Ali[3], Volodymyr Lytvynenko

1 - University of Jan Evangelista Purkinje, Usti nad Labem, Czech Republic
2 - National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine
3 - Kherson National Technical University, Kherson, Ukraine

sergii.babichev@ujep. czvvo7@ukr.net
taifmohamedali@gmail.com, immun56@gmail.com

Досліджена можливість застосування бікластерного аналізу в системах кластеризації об'єктів складної біологічної природи. Бікластерізація проводилася за алгоритмом BCCC пакету "biclust" програмного середовища R, кластеризація об'єктів в бікластерах проводилася за алгоритмом SOTA. Критерієм оцінки якості кластеризації було відношення середньої евклідової відстані між об'єктами в різних кластерах і в окремих кластерах відповідно.

Ключові слова. Бікластеризація, кластерний аналіз, мікромасив ДНК, алгоритм SOTA.

In article studied the possibility using of bicluster analysis in clustering systems of complex biological nature objects. Biclustering was performed by use the algorithm BCCC of package "biclust" of software environment R, clustering of objects in each bicluster was performed by use the algorithm SOTA. The ratio of the average Euclidean distance between objects in different clusters and individual cluster accordingly were used as the criterion for an estimation of quality clustering.

Keywords. Biclustering, cluster analysis, DNA microarray, SOTA algorithm.

Исследована возможность применения бикластерного анализа в системах кластеризации объектов сложной биологической природы. Бикластеризация проводилась с использованием алгоритма BCCC пакета “biclust” програмной среды R, кластеризация объектов в бикластерах производилась с использованием алгоритма SOTA. В качестве критерия оценки качества кластеризации использовалось отношение среднего евклидового расстояния между объектами в разных кластерах и в отдельных кластерах соответственно.

Ключевые слова. Бикластеризация, кластерный анализ, микромассив ДНК, алгоритм SOTA.

Problem statement. In the process of data analysis of complex biological nature when solving problems of classification and clustering of researched objects arises the problem of finding an optimal model, allowing to obtain the required accuracy of objects partitioning into classes or clusters. One of such tasks is the analysis of gene expression profiles of DNA microarray in order to allocate the genes whose expression corresponds to the type of disease and the nature of the object and disease progression. Model selection is determined by the view of the objective function and the structure of the test data. The complexity of the problem to be solved is determined by:

- the nature studied data, of which feature is the high level and the specificity of the noise components arising during microarray production and reading information from it;

– high dimensional feature space of studied genes.

Using the factor analysis partially solves the problem of reducing the features dimension, however, in the process of the data transformation takes place a partial loss and the distortion of the initial information that has a direct impact on the accuracy of the problem being solved. The use of bicluster analysis retains object-attributive data structure, but the dimension of the feature space in obtained of subsets of researched objects much smaller than the dimensions of the original data that allows the clustering them in real time. As a result of parallel processing biclusters is the obtained a set of solutions, that can both complement each other, and engage in certain contradictions. Thus, there is an actual problem the searching of complex criteria or groups of criteria, which the extremum on different subsets of researched objects will enhance the objectivity of the clustering of objects of complex biological nature.

Analysis of publications on the study subject. The biclusterization questions of gene expression data are considered in [1, 2]. The authors analyzed various algorithms and highlighted their advantages and disadvantages. In [3] was conducted a comparative analysis of different biclusterization algorithms for the analysis of gene expression profiles. In [4] are presented a study on the use of spectral biclusterization technique for the analysis of gene expression data and by the example of simulated data displayed the distribution diagram of objects and characteristics of their grouping in different biclusters. In [5] the author describes the possibility of solving the problem of clusterization by inductive methods for modeling of complex systems using the target criterion in order to determine the optimal partition of objects into clusters. However, it should be noted that, despite the apparent success of this subject area, the problem of partitioning the nature of complex biological objects into classes or clusters at the present time has no unique solution.

As unsolved aspects of the problem is the absence of a methodology clusterization of complex biological nature objects with parallel using a subset of features, obtained by biclusterization of the researched objects initial set.

The aim of the article is the study of applying possibility of bicluster analysis in systems of clusterization objects of complex biological nature with the use of inductive approach for modeling of complex systems.

The presentation the basic material. The matrix of gene expression profiles can be represented as follows:

$$A=\left\{a_{ij}\right\}, i=1...n, j=1...m \quad (1)$$

,

wherein a_{ij} – the expression level of i-th gene under the condition j-th, n - number of the genes are studied, m – number of conditions under which the process of hybridization of the corresponding gene was carried out.

Let $X=\left\{x_1,...,x_n\right\}$ is the vector representing the set of rows of the matrix, and $Y=\left\{y_1,...,y_m\right\}$ – set of columns. Then the matrix A can be represented as $A=(X,Y)$. As a result of biclusterization the matrix A is broken down into sub-matrix $A_{IJ}=\left(I,J\right)$, where $I=\left\{i_1,...,i_k\right\}\subseteq X$ – is the subset of rows, and $J=\left\{j_1,...,j_s\right\}\subseteq Y$ –

is the subset of the bicluster columns. Under the bicluster we understand a set of objects, the feature vector and the conditions which have the mutual functional similarity. By analogy with (1) the matrix of genes profiles in bicluster has the form:

$$B = \{b_{ij}\}, i = 1...k, j = 1...s \tag{2}$$

The difference between the classical clustering from biclustering is shown in Fig. 1. The classical clustering algorithms divide the input data into clusters considering the full feature vector that characterizes the object. However, in the case of high-dimensional feature space, the different characteristics inherent one gene may contradict each other. Fig. 1a shows the conditional a partition of set objects into 5 clusters. At the same to each cluster corresponds a full vector feature space. With the help of biclustering are allocated objects, features of which correspond to similar hybridization conditions (Fig.1b). In this case, the length of the feature vector dimension is substantially smaller than the feature space of the original matrix. However, at the same time there is a problem of bicluster intersection.

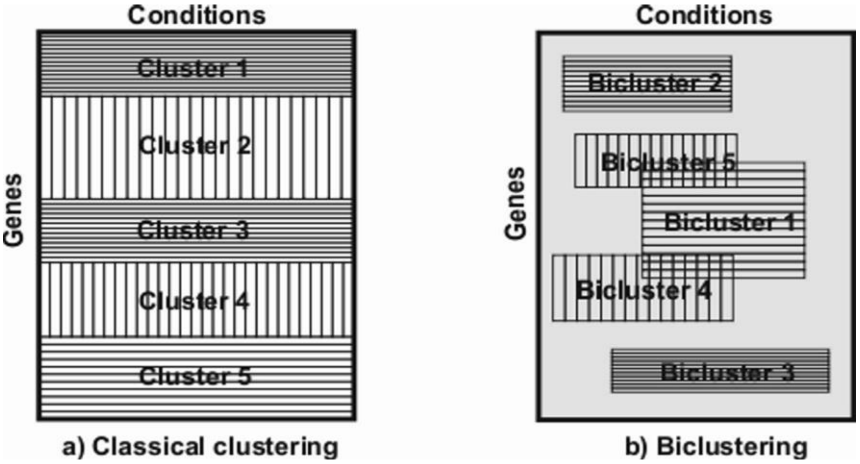


Fig.1. Models of cluster analysis: a) the classical clustering of genes;
b) biclusterization on genes and environment conditions.

The model classification of the types of clusters is presented in [6]. Let δ – is a constant of bicluster B and $\beta_i (1 \le i < k)$ and $\beta_j (1 \le j < s)$ are the constant values for the gene or condition, respectively. The average values of gene expression in the *i*-th row b_{iJ}, *j*-th column b_{Ij}, and in bicluster b_{IJ} defined as follows:

$$b_{iJ} = \frac{1}{s}\sum_{j=1}^{s} b_{ij} \tag{3}$$

$$b_{Ij} = \frac{1}{k}\sum_{i=1}^{k} b_{ij} \tag{4}$$

$$b_{IJ} = \frac{1}{k}\sum_{i=1}^{k} b_{iJ} = \frac{1}{s}\sum_{j=1}^{s} b_{Ij} \tag{5}$$

Bicluster with constant values represents a matrix *(I, J)*, in which all elements are equal:

$$b_{ij} = \delta \tag{6}$$

When developing of selection algorithms of one type of cluster with constant values, as the primary criterion for the algorithm stopping is used the dispersion of bicluster gene features. For one bicluster:

$$VAR\left(I, J\right) = \sum\nolimits_{i \in I, j \in J} \left(b_{ij} - b_{IJ}\right)^2 \tag{7}$$

In the presence of the *K* biclusters the clustering quality criterion is calculated as the total dispersion on all biclusters:

$$VAR\left(I, J\right)_{\acute{}K} = \sum_{k\acute{}1}^{K} \left(\sum\nolimits_{i \in I, j \in J} \left(b_{ij} - b_{IJ}\right)^2 \right) \tag{8}$$

Bicluster with constant values for the rows and columns represents a matrix in which the element values are calculated as follows:

$$\begin{gathered} b_{ij} = \delta + \beta_i, \;\; b_{ij} = \delta \times \beta_i, \\ b_{ij} = \delta + \beta_j, \;\; b_{ij} = \delta \times \beta_j \end{gathered} \tag{9}$$

In [7-9] are presented the algorithms of this type allocation of biclusters. At the initial stage of the algorithms the data are normalized by rows or columns of the matrix. The normalization process allows the identification of rows and columns with the same characteristic values. As a metric is used the Euclidean distance.

The cluster elements with coherent value is calculated as follows:

$$\begin{gathered} b_{ij} = \delta + \beta_i + \beta_j, \\ b_{ij} = \delta \times \beta_i \times \beta_j, \end{gathered} \tag{10}$$

One of the first algorithms to allocate the bicluster with coherent values is described in [10]. As a closeness measure of rows and columns in bicluster authors used the standard error:

$$H\left(I, J\right) = \frac{1}{k \times s} \sum_{i=1}^{k} \sum_{j=1}^{s} \left(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ}\right)^2 \tag{11}$$

At the initial stage it is specified the delta-parameter$\delta \geq 0$, which is a maximum value of RMS of bicluster. The duration of the algorithm operation is limited by the condition of $H\left(I, J\right) \leq \delta$. The delta-parameter determines the structure of biclusters and their amount.

In our case, the objects clustering in each bicluster carried out using an algorithm SOTA (Self-Organizing Tree Algorithm) [11, 12]. The SOTA algorithm generates a binary topological tree in accordance with the principles of growing the cell structure of the algorithm Fritzke. The algorithm SOTA takes into account the heterogeneity of Fritzke distribution facilities in the feature space.

During the SOTA-algorithm run, the sequence of binary tree nodes is adapted to the characteristics of the feature space of the input data set. Herewith, the number of output nodes in the model fitting process is determined by varying of the input data of feature space. As a measure of the vectors similarity was used the correlation or Euclidean metric, determined in accordance with formulas:

$$d_{12} = \sqrt{\sum_{i=1}^{n} \left(b_{1i} - b_{2i}\right)^2} \tag{12}$$

$$d_{12} = (1 - r) = 1 - \frac{\sum_{i=1}^{n}\left(\left(b_{1i} - \bar{b}\right)\cdot\left(b_{2i} - \bar{b}_2\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(b_{1i} - \bar{b}_1\right)^2 \cdot \sum_{i=1}^{n}\left(b_{2i} - \bar{b}_2\right)^2}}, \qquad (13)$$

where r – Pearson's correlation coefficient, $\bar{b}$ – the average value of the corresponding vector characteristics.

In the initial state, the system is composed of two cell units pooled by the external root node, i.e. it has the structure of a binary tree (Fig. 2a). Each node is characterized by a vector of features, the number of which is equal to the dimension of the studied genes feature space. The value of each feature in a column of the vector defines the conditions under which the measurement of expression of the corresponding gene.

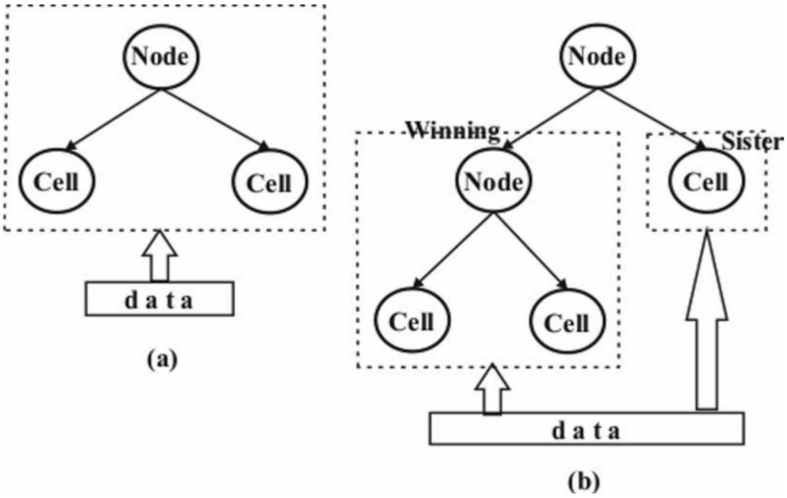


Fig.2. The cell formation structure by the algorithm SOTA:
a) the initial state of the system; b) the status of the system after one cycle.

The algorithm SOTA consists of the following stages:

1. *Initialization.* For features of the root node vectors and cells are assigned the weights which values are equal to the average features value of the all studied data columns. The length of weights vector is equal to the dimension of studied data feature space.

2. *Adaptation.* During algorithm operation the input of the all external cells is sequentially supplied by feature vector of studied objects. Then the degree of closeness of this vector with weights the cells is calculated by formulas (12) or (13). In accordance with the principle of "winner takes all" is allocated the cell-winner, which vector of weights has the smallest distance from the investigated gene profiles vector. The weights of the cell-winner and its vicinity are adjusted in accordance with the formula:

$$C_i(\tau + 1) = C_i(\tau) + \eta \cdot \left(P_j - C_i(\tau)\right), \qquad (14)$$

where $C_i(\tau)$ and $C_i(\tau + 1)$ – are the cells i weight vectors in step τ and τ*+1*, respectively, P_j – is the profiles vector *j*-th gene on the input system, η – is the parameter that determines the adjusting step of the cell-winner weights which at iteration *t* is defined as [11]:

$$\eta_t = \alpha \cdot \frac{1-t}{n} \cdot \left(1 - b\tau\right) \tag{15}$$

and where *t* – is the total number of objects, *n* – is the maximum number of studied objects, τ – is the number of operations per cycle, *b* – coefficient that determines the rate of change of the parameter η, α – is the parameter determined by empirically proceeding from a condition: $\alpha_w > \alpha_m > \alpha_s$ where α_w, α_m and α_s – are the coefficients for adjusting weights of the cell-winner that binds the node with its neighbor cell respectively. The values of *α* and *b* for neighboring cells and the node are selected empirically.

3. *The convergence of the algorithm and the network formation.* To determine the clusterization tree structure, the variation coefficient of each cell as the arithmetic average value between the values of the cell weights and gene expression profiles values in this cell is calculated as:

$$R_i = \frac{\sum_{i=1}^{k} d_{P_k C_i}}{k} \tag{16}$$

The total value of the variation coefficient is defined as the sum of the variation coefficients for all external cells:

$$\varepsilon_t = \sum_{i=1}^{s} R_i \tag{17}$$

The assessment criterion of the algorithm convergence is the relative change in the total coefficient of variability:

$$\left| \frac{\varepsilon_t - \varepsilon_{t-1}}{\varepsilon_{t-1}} \right| < E \quad , \tag{18}$$

where *E* is the threshold tresholding factor. The cycle ends if the condition is (18).

The further growth of the tree begins with the cell having the largest value of the coefficient of variation. This cell is divided into two parts and becomes a node (Fig. 2b). The weighting values of daughter cells and a node are identical to each other. The growth of network is finishes when the total value of the variation coefficient reaches a certain threshold value. At the zero value the threshold coefficient the number of clusters equal to the number of studied objects.

The simulation of clustering process with subsequent analysis of the selected genes was carried out using the data for lung cancer patients GEOD-E-68 571 Array Express database [13]. The data included the gene expression profiles of 96 patients, of which 10 were healthy (Norm), and 86 patients have been divided according to the degree of the disease into three groups: 24 patients have a good condition (Well), 41 patients have a mild condition (Moderate-Md) and 21 patients have a poor condition (Poor). Each object was characterized by 7129 signs. The delta-parameter empirically was selected so that in the first bicluster the all studied objects are contained.

Biclustering simulation process was carried out using an algorithm BCCC of package "biclust" of software environment R [10]. In the process of algorithm operation were allocated the biclusters in which the mean square error did not exceed a priori specified parameter delta of δ. During algorithm operation has been allocated 100 biclusters. The dependence graph of the conditions number of hybridization process from the amount of genes in this biclusters is shown in Fig. 3.

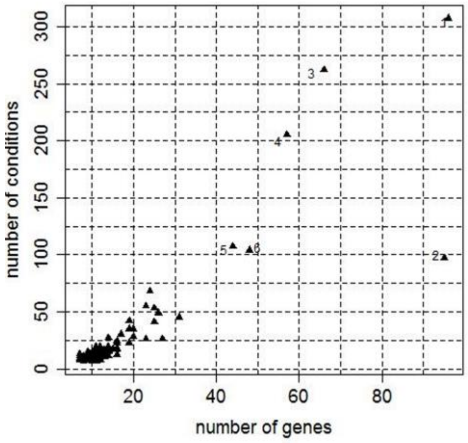


Fig. 3. The graph of genes distribution and hybridization conditions depending on the number of clusters

An analysis of Fig. 3 allows drawing a conclusion that the most informative for further analysis are the first six biclusters, because they contain the greatest number of studied genes. The results of cluster analysis of objects from first bicluster with using of SOTA-algorithm with allocation the four clusters are shown in Fig. 4. The similar distributions of objects were obtained for other biclusters. An analysis of Fig. 4 leads to the conclusion about inexpedience of partitioning the objects at each step more than 2 clusters, since the bulk of the objects contained in the first two clusters. At the same time, the small part of the remaining objects in the cluster can be treated as emissions.

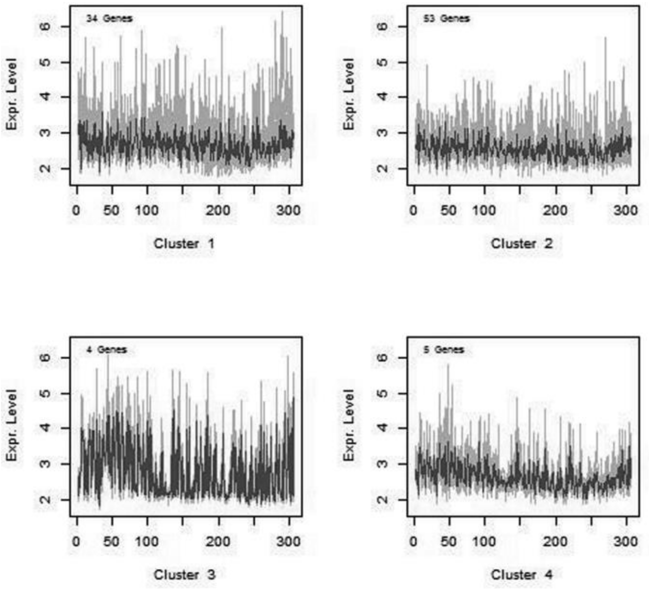


Fig. 4. The results of cluster analysis the objects in the first bicluster.

The estimation of the optimal number of steps was made on the basis of the coefficient which was defined as the ratio of the average Euclidean distance between objects of different clusters to the average Euclidean distance between objects within the corresponding cluster:

$$k = \frac{\bar{d}_{between}}{\bar{d}_{inside}}, \qquad (19)$$

where $\bar{d}_{inside}$ and $\bar{d}_{between}$ – are the average Euclidean distances between all pairs of objects in a same cluster and in the different clusters respectively:

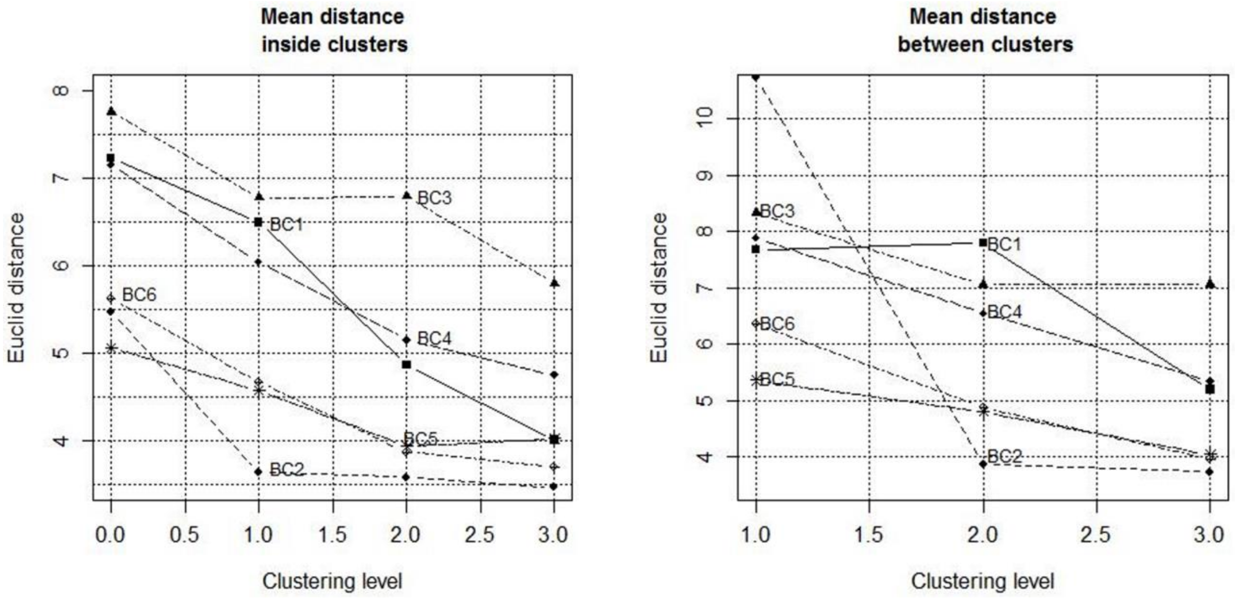
$$\bar{d}_{inside} = \left(\frac{1}{n(n-1)} \sum_{s=1}^{n} \sum_{t=1}^{n} \sum_{i=1}^{m} (b_{si} - b_{ti})^2 \right)^{\frac{1}{2}}, \qquad (20)$$

here: n – the number of objects in a cluster, m – the dimension of feature space of studied objects. And:

$$\bar{d}_{between} = \left(\frac{1}{n \cdot k} \sum_{s=1}^{n} \sum_{t=1}^{k} \sum_{i=1}^{m} (b_{si} - b_{ti})^2 \right)^{\frac{1}{2}}, \qquad (21)$$

where n and k – the number of objects in the first and second clusters respectively.

Fig. 5 shows diagrams of dependence the average Euclidean distances within clusters (Fig. 5a) and between clusters (Fig. 5b) on the level of clusterization. The zero level corresponds to the initial bicluster.



(a) (b)

Fig. 5. Plots of dependence the Euclidean distance
from the level of clusterization: a) within the cluster; b) between clusters.

Fig. 6 illustrates the dependence graph of the ratio distances of the coefficient from the level of clusterization. An analysis of drawings allows us to conclude that for the first cluster the optimal level is the second clusterization level, since the average value of the Euclidean distance within clusters decreases (Fig. 5a), and between clusters increases (Fig. 5b). As a result, on the basis of analysis of the first 6 biclusters can be identified 14 clusters: 4 – from the first bicluster and 2 from the remaining five. The increasing the number biclusters will allow to receive a greater of clusters number, but increases in this case the complexity of data processing.

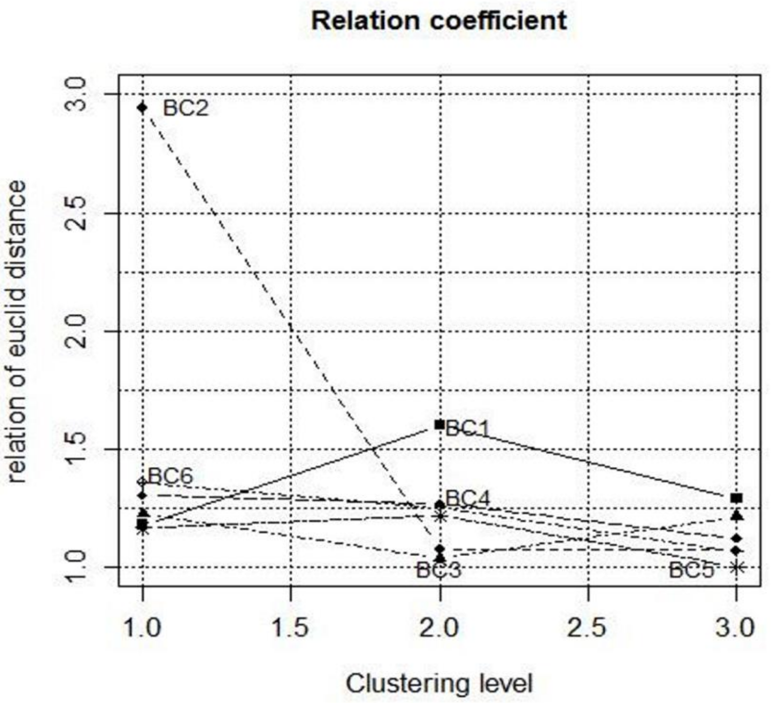


Fig. 6. Plot of dependence of ratio coefficients of Euclidean distance from the clusterization level.

Conclusions. Conducted research have shown the prospects of bicluster analysis application to the data processing of complex biological nature having high levels of noise component and high dimension of feature space. Biclustering of the initial objects set allows to allocate a subset of objects, the feature space of which was formed under similar conditions. The biclusters with constant values contain the noise component and their exclusion from further analysis is equal to the process of filtering data. The interest for further analysis are the biclusters with coherent characteristic features, since they have the basic information about the object. Prospects for further research of the authors is to create the new information technology of parallel data processing in allocated clusters in order to regroup the objects in accordance with the required criteria for the formation of clusters based on the inductive modeling of complex systems methodology.

Bibliography

1. Pontes B., Giráldez R., Aguilar-Ruiz J.S. Biclustering on expression data: A review // Journal of Biomedical Informatics, 2015.– №57.– 163-180.

2. Kaiser S. Biclustering: Methods, Software and Application. – Minchin, 2011. – 163 p.

3. Eren K., Deveci M., Kucuktunc O.¸ Catalyurek U.V. A comparative analysis of biclustering algorithms for gene expression data // Briefings in Bioinformatics, 2012. – V. 14, №3. – P. 279-292.

4. Kluger Y., Basry R., Chang J.T., Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions // Genome Resources, 2003. – №13(4). – P. 703-716.

5. V. Osypenko. Two approaches to solving the problem of clustering in the broad sense from the standpoint of inductive modeling // Power and Automation, 2014. - №1. - P.83-97. [In Ukraine].

6. Mukhopadhyay A., Maulik U., Bandyopadhyay S. On biclustering of gene expression data // Current Bioinformatics, 2010. – №5.– p. 204-216.

7. Califano A., Stolovitzky G., Tu Y. Analysis of gene expression microarrays for phenotype classification // In Proceedings of the International Conference on Computational Molecular Biology, 2000.– p. 75-85.

8. Getz G., Levine E., Domany E. Coupled two-way clustering analysis of gene microarray data // In Proceedings of the Natural Academy of Sciences USA, 2000.– V.27(22).– p. 12079-12084.

9. Sheng Q., Moreau Y., De Moor. B. Biclustering microarray data by Gibbs sampling // Bioinformatics, 2003.– V.19(2).– P. 196-205.

10. Cheng Y., Church G.M. Biclustering of expression data // Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB’00), 2000.– P. 93-103.

11. Dorazo J., Carazo J.M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree // Journal of Molecular Evolution, 1997. – №44(2).– P. 226-259.

12. Fritzke B. Growing Cell Structures A Self-Organizing Network for Unsupervised and Supervised Learning // Neural Networks, 1994.– V.7, №9.– P. 1441-1460.

13. Beer D.G., Kardia S.L., 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., Iannettoni M.D., Orringer M.B., Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma // Nature Medicine, 2002. – № 8(8). – P. 816-824.