Analysis of Short Time Series in Gene Expression Tasks

Arnis Kirshners, *Riga Technical University*, Arkady Borisov, *Riga Technical University*

Abstract – **The article analyzes various clustering approaches that are used in gene expression tasks. The chosen approaches are portrayed and examined from the viewpoint of use of data mining clustering algorithms. The article provides a short description of working principles and characteristics of the examined methods and algorithms and the data sets used in the experiments. The article presents results of the experiments that are directly connected to the use of clustering algorithms in processing of short time series in bioinformatics tasks, solving gene expression problems, as well as provides conclusions and evaluations of each used approach. An analysis of future possibilities to build a new method that is based on data mining approaches and principles but solves bioinformatics tasks that are associated with processing of short time series and the achieved results are interpreted in a way that is easy to perceive for bioinformatics experts, is presented.**

Keywords – **gene expression, data mining, short time series, clusterization algorithms**

I. INTRODUCTION

Usually for data clustering to solve various tasks in bioinformatics one of the following methods is used: clustering with neural networks, time series clustering or fuzzy clustering Chaiboonchoe et al. [1].

To visualize data and to reduce dimensionality, neural networks are used with non-linear clustering method that is also known as self-organizing maps (SOM). Neural networks are often used to analyze gene expressions Golub et al. [2], where clustering is used for example in differentiate between different types of leukemia without prior knowledge about the classes of these types. Based on the acquired knowledge the membership of a new case of leukemia to one of the classes is predicted.

Time series are used to determine mechanisms and dependencies that emerge when various diseases progress in a period of time Ernst et al. [3]. To investigate these processes a tool called STEM was created (Short Time-series Expression Miner). It is designed to analyze and process short time series of expressions Wang et al. [4].

In bioinformatics more than 80 % of the studied time series data sets are considered short, because the number of examined genes is very large (it can reach even tens of thousands), but the number of observations of these genes in time (periods) is small. Mostly the number of these periods in bioinformatics is 8 or less Ernst et al. [5], see Figure 1.

In experiments that use short time series as the processed data source, e.g. examining gene expressions that are influenced by environmental or chemical factors that create changes in their structure, these observations are recorded in a short period of time for example every 0, ½, 4 and 16 hours.

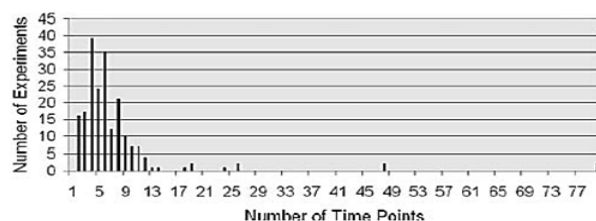The structure of the obtained curve is used as a basis for further gene expression analysis.

Fig. 1. Distribution of lengths of times series in the Stanford Microarray Database Ernst et al. [5]

The goal of this article is to summarize the methods and algorithms that are used by various authors for clustering short time series and to analyze clustering approaches that are used in bioinformatics to solve gene expression tasks.

II. BACKGROUND

Genomics studies methods of molecular cartography and sequencing as well as description and analysis of full genomes of organisms. Genome is a full set of genetic material that is stored in the set of chromosomes. The analysis of full genomes gives insight into global organization, expression, regulation and evolution of the hereditary material.

Gene expression that uses the gene as a synthesis plan of a particular protein is called identification of potential of a gene. Images of gene expressions enable scientists to find approaches to determine the biological roles of particular genes. All functions of cells, tissues and organs are managed by differential gene expressions.

The analysis of gene expressions is used to explore functions of genes. By acquiring information about genes that are expressed in healthy and diseased tissues the scientists can define a set of proteins that is essential for normal functioning of a gene. The obtained information is also helpful in determining deviations from the normal set that can point to disease agents. Such information opens up new possibilities in creating new diagnostics tools and techniques of various diseases and developing new medicines that can actively affect defective genes and proteins Ignacimuthu [6].

III. METHODS

The article examines methods proposed by some authors for short time series processing in solving gene expression tasks that are associated with the use of clustering approaches. The chosen methods are examined from the point of view of data processing techniques and algorithms used in data mining.

Cluster analysis of gene expression dynamics Ramoni et al. [7] acquires profiles that are estimations of average values of clusters that are obtained using autoregressive equations and agglomerative clustering Tan et al. [8]. Cluster analysis and display of genome-wide expression patterns Eisen et al. [9] method uses search of similarities between genes based on Pearson correlation coefficient creating matrixes that provide profiles using hierarchical clustering. But clustering of short time series in gene expression data Ernst et al. [5], studies possibilities of profile acquisition from data sets using clustering and forming groups using the acquired profiles.

A. *Cluster analysis of gene expression dynamics*

The proposed method is based on Bayesian method that chooses a model with the maximum posterior probability. The dynamics of gene expression are assessed based on autoregressive equations and agglomerative clustering procedure. In this approach the time series S, see Equation (1), is considered a stochastic process with unknown length.

$$S = \{S_1, S_2 \dots S_n\}, \tag{1}$$

The aggregation of time series into groups occurs iteratively by repeatedly adding a time series to a cluster so that each of the clusters would contain time series that are obtained during the same process. This clustering method consists of two steps:

1. Time series with stochastic descriptions are clustered acquiring probability assessments that determine approximation of the processes that generated the data using autoregressive models;
2. Heuristic search procedure.

The heuristic search procedure is based on estimation of mutual similarity of two time series $S_i=\{y_{i1}, \dots, y_{in}\}$ and $S_j=\{y_{j1}, \dots, y_{jn}\}$. In this research Euclidean distance that is calculated as shown in the Equation (2) is chosen as the measure of similarity. First the similarity of pairs of time series is calculated, then a model S_{n-1} is made that merges the

$$d(S_i, S_j) = \sqrt{\sum_{t=1}^{n}(x_{it} - y_{jt})^2}, \tag{2}$$

closest time series into one cluster. If the Equation (3) confirms the statement, these two time series are considered to be merged into one cluster.

$$f(S|S_{n-1}) > F(S|S_n), \tag{3}$$

The profile of the cluster is calculated based on average values of the considered time series. The procedure is repeated for the new data set that consists of *n-1* time series, including the new cluster profile. If Expression (3) is not confirmed, the process is repeated using a pair of time series with lower similarity until the Expression (3) is confirmed. If the confirmation of the Expression (3) is not found, the procedure is interrupted. The authors of this method emphasize that a merger of two clusters is accepted based on posterior probabilities of the model and the similarity measure is only

used to avoid the risk of getting into local minimum Ramoni et al. [7].

The authors also point out that the method does not contradict the Standard hierarchical clustering method Tan et al. [8] and works of other authors Golub et al. [2], who investigate clustering of gene expression data.

This method was applied to 517 gene expressions that were described by time series with length of 13 periods. Series of experiments were conducted to solve the problem that is associated with time dependency between periods using the following measures of distance: (i) Euclidean distance; (ii) Correlation; (iii) Delayed correlation; (iv) Kullback-Leiber distance. The best clustering results were shown by Euclidean distance Ramoni et al. [7].

B. *Cluster analysis and display of genome-wide expression patterns*

This method uses standard statistics algorithms to arrange genes according to pattern of gene expression. Clustered gene expression data are grouped according to a similarity function and a resemblance to human gene data is found. Initially standard deviation from G is calculated using Equation (4):

$$\Phi_G = \sqrt{\sum_{i=1,N} \frac{(G_i - G_{offset})}{N}}, \tag{4}$$

where

Φ_G – standard deviation;

G_i – primary genes of the data in state *i*;

G_{offset} – set to the mean of observations on *G*;

N – number of observation series.

G_{offset} in the given examples is equal to 0 that corresponds to a fluorescence ratio 1,0. Then according to Equation (5) $S(X,Y)$ is calculated; it is equal to Pearson correlation coefficient for observations *(X,Y)*:

$$S(X,Y) = \frac{1}{N}\sum_{i=1,N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \tag{5}$$

where

X,Y – any two genes;

S(X,Y) – Pearson correlation coefficient;

N – number of observation series.

After determining gene similarity using hierarchical clustering that is based on average linkage method of Sokal and Michener Sokal et al. [10] correlation matrixes are clustered. The algorithm builds a dendrogram that combines all elements into one tree. For any number *n* of genes in the set a diagonal similarity matrix is calculated according to metric that was described previously and that contains similarity measures for all pairs of genes. Matrix search provides the most appropriate pair of genes that is added to the matrix as a new node. Then the search is repeated *n-1* times until all nodes are mutually connected, that is, a dendrogram is built.

To improve the interpretability of the results the authors used fluorescence coefficient. Every cell is colored according to the value of this coefficient. Genes that did not change the

value of the coefficient (0) are colored black, genes that have a negative value of the coefficient are colored green but genes with a positive value of the coefficient are colored red Eisen et al. [9].

In this research two types of data sets were used – one depicts a canonical model of the growth response in human cells Iyer et al. [11], but the other outlines experiments with yeast.

The experiments with human cells showed the stimulation of the cells caused by serum that was recorded after various periods of time: 0, ¼, ½, 1, 2, 3, 4, 8, 12, 16, 20 and 24 hours. For this purpose a cDNA microarray with nearly 8600 human genes was used. In the experiments with yeast 2467 genes were divided into groups according to profiles that contained similar genes using clustering Eisen et al. [9].

C. *Clustering short time series gene expression data*

The method proposed by authors is based on initially adding genes to profiles and then dividing these profiles based on relevance. The number of profiles P is calculated using Equation (6):

$$P = (2c + 1)^{n-1}, \qquad (6)$$

where

c – controls the amount of changes in a gene that can be displayed between consecutive moments of time;

n – number of time periods in a data set.

For example, if the number of time periods in a data set is $n=6$ and $c=2$, then $5^5=3125$ profiles are obtained, which is a very large number. Therefore, the number of profiles should be lessened. From a set P, containing all possible profiles, a subset R, $R \subset P$ should be selected, following the condition that a minimal distance between profiles in R should be maximized using Equation (7).

$$\max_{R \subset P} \min_{p_1, p_2 \in R} d(p_1, p_2), \qquad (7)$$

where d– distance measure; p_1, p_2 – time series for which d is calculated.

The set R contains profiles that were selected at this moment. The next profile that will be added to R will be profile p_1, which is chosen by maximizing Equation (8):

$$\max_{p_1 \in (P \backslash R)} \min_{p_2 \in R} d(p_1, p_2). \qquad (8)$$

In every iteration the method chooses a profile that is furthest from all previously chosen R profiles. This procedure is repeated until all R profiles are selected and the obtained result is returned to the data set.

Then significant profiles are grouped, increasing the size of cluster C_i around each statistically significant profile p_i. Initially $C_i = \{p_i\}$, then the method searches for a profile p_j that is the closest to profile p_i and does not belong to C_i. If the condition $d(p_j, p_k) \leq \delta$ is true for all profiles $p_k \in C_i$, then the profile p_j is merged with C_i and the process is repeated until the algorithm stops and C_i is declared cluster p_i. When all clusters of significant profiles are obtained, the algorithm

chooses a cluster with the largest number of genes by calculating the number of genes in each profile cluster. Then all profiles of this cluster are deleted and the previously described process is repeated. The algorithm stops when all profiles are added to clusters. Execution time of the algorithm depends on the number of significant profiles that usually is not large Ernst et al. [5].

A data set that was used for experiments consisted of 5000 genes that were described by time series with period length 5 (time intervals – 0, ½, 3, 6, and 12 hours). 50 profiles were used that had the maximum change between consecutive moments of time where $c=2$. A visualization of the obtained clustering results that is generated by a specialized tool for time series processing called STEM is shown in Figure 2. These profiles that were obtained using clustering show that from 50 profiles 10 were considered significant (colored darker in Figure 2). This allows drawing biological conclusions based on the revealed significant profiles Ernst et al. [5].
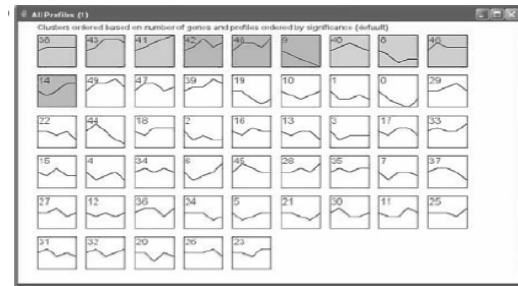

Fig. 2. Clustering results with the STEM tool Ernst et al. [5]

Only the acquired significant profiles are scientifically relevant and provide new knowledge that enables drawing biological conclusions. The essence of these conclusions and their right interpretation can produce new added value.

IV. CONCLUSIONS

The proposed evaluation of the applied methods proves that cluster analysis of gene expression dynamics that proposes using gene clustering based on their dynamics is not applicable to clustering of short time series. This method is based on regression and clusters genes that have dynamics that can be described using autoregressive equations meaning that this method is effective only when working with long time series. For example, if regression is applied to time series with expressions in five moments of time, then only last three can be used and the first two expressions cannot be regressed. Such a situation can lead to overfitting and an imperfect cluster distribution.

The most relevant knowledge gained from this method was improved and used in the future to develop noise processing method for expression data and also to achieve better functionality and find more meaningful clusters based on gene ontology Wang et al. [4], Wang et al. [12]. By introducing the ontology gene expression profile was supplemented by noise measures in the stochastically dynamic model. This allows determining distribution of noise by analyzing microarrays of

time series and calculating relationships of short time series that match the parameter evaluation requirements of multi-criteria traditional methods Wang et al. [13].

By analyzing the method of cluster analysis and display of genome-wide expression patterns and the data sets used in experiments one can conclude that these data sets were especially useful for division of genes into functional categories. Hierarchical clustering was used for splitting the data using links between times series that were defined by estimating similarity measure based on correlation coefficient. Positive features of the proposed approach are related to the use of colors for description of gene correlations. The use of colors allows visually determining the belonging of genes to one group or another that relieves interpretation of results.

Another approach that was examined was clustering of short time series in gene expression data that uses aggregating genes based on probability assessments of chosen sub-profiles that were generated independently from short time series data that were used in experiments. The proposed approach also provides evaluation of the created method where the authors chose 50 genes (1%) from the initial data set and generated the remaining 4850 using the previously examined method. The results of the experiments show that these three profiles above diagonal line (see Figure 3) were identified by the algorithms as noise Ernst et al. [5]. This means that the proposed algorithm can interpret biological results that generate biologically valuable data.
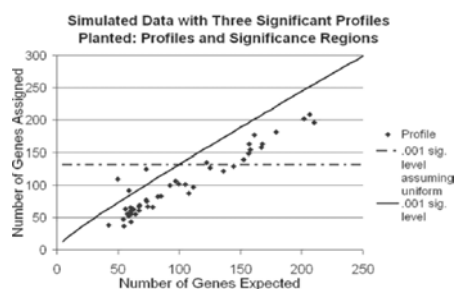

Fig. 3. Similar plot for algorithm correctly identified all three planted profiles, even though each was planted with only 1% of the genes Ernst et al. [5]

The clustering of gene expression data can be described as a process that creates associations Androulakis et al. [14] between profiles.

In general, one can conclude that this field is at the very beginning of research because most of the researches that deal with processing of short time series use improved methods that were initially meant to work with long time series. Unfortunately, modules of these methods are based on statistical criteria that limit their biological relevance when used to solve bioinformatics tasks.

Data mining also solves tasks that are related to processing short time series of demand, e.g. analysis of e-services demands Kirshners et al. [15] that use combinations of methods in the research. These combinations include analysis of time series and data mining methods providing new knowledge as a result. This allows giving recommendations for e-service introduction based on this knowledge. Time series represent frequency of service demands in a defined

period of time. Bioinformatics have similar situation because not all achieved results are interpretable. Therefore bioinformatics experts work closely with experts from industry who can interpret the obtained results.

Data mining also deals with different type of problems that are related to processing of short time series. Traditional forecasting systems Montgomery et al. [16] search for patterns in time series, e.g. fluctuations in exchange rates Flores et al. [17], but there are tasks where the life cycle of a time series is short Kirshners et al. [18] and determining regularities in the functioning of a process is almost impossible. One of these tasks is the task of analyzing life cycle of a product Thomassey et al. [19], e.g. the demand of textile goods. It defines a task that analyzes the behavior of short time series with a large number of demands and a small number of periods. To solve the described task, a modified k-means algorithm was used that determined the optimal number of clusters in a data set and created profiles based on results of clustering Kirshners et al. [20]. The features of this method show a similarity with the approaches described in the article. Therefore approaches used in this method could also be used for solving bioinformatics tasks.

The examined results of the inspected methods allow one to conclude what should be taken into account when developing a more efficient clustering method for short time series.

V. FURTHER RESEARCH

In the future when creating a new time series clustering method, more attention should be paid to processes of data preprocessing and evaluation of obtained results. Data mining methods require preprocessing and normalization of noisy time series in the data preprocessing process. The process of normalization should be inspected in more detail because there is a chance that different types of time series data require different data normalization approaches. More attention should also be paid to determining the number of clusters in the data set because the methods described in this article did not analyze this process. It is necessary to create approaches that would evaluate mistakes made in the process of clustering because it is hard to validate adequacy of the results of the created method without the estimation of errors. Interpretation of results is also very important because results of clustering using data mining methods do not always show the real cluster distribution as seen by bioinformatics, therefore the results should be accurately interpreted. To gain such results the created method should be evaluated using a data set with existing results and known distribution of classes or clusters and the clustering results should be evaluated during the process of clustering; only this evaluation can indicate compliance of the method with the analyzed data.

By analyzing the principles used for profile creation in the examined methods conclusions can be drawn that the use of these methods effectively helps combining similar time series in one profile. This approach reduces the number of analyzed objects and the results are easily visually interpretable from the point of view of data mining as well as bioinformatics. Only a combination of all listed methods and approaches

would ensure creation of an efficient short time series clustering method for bioinformatics tasks.

It is important to remember that not all results obtained using data mining methods, even with very good results are equally efficiently interpretable in the field of bioinformatics. Therefore differentiation of the obtained results between data mining and bioinformatics is necessary. Only if experts of both fields have acknowledged the interpretability of the results the created method can be accepted as adequate for solving this type of tasks.

REFERENCES

[1] **A. Chaiboonchoe**, **S. Samarasinghe**, **D. Kulasiri**, *Using emergent clustering methods to analyse short time series gene expression data from childhood leukemia treated with glucocorticoids. 18th World IMACS/MODSIM Congress*, July 13-17, 2009, Cairns, Australia, pp. 741-747.

[2] **T. R. Golub**, **D. K. Slonim**, **P. Tamayo**, **C. Huard**, **M. Gaasenbeek**, **J. P. Mesirov**, **H. Coller**, **M. L. Loh**, **J. R. Downing**, **M. A. Caligiuri**, **C. D. Bloomfield**, **E. S. Lander**, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286 (5439), 1999, pp. 531-537.

[3] **J. Ernst**, **Z. Bar-Joseph**, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, 2006, 7:191.

[4] **X. Wang**, **M. Wu**, **Z. Li**, **C. Chan**, "Short time-series microarray analysis: methods and challenges," *BMC Systems Biology*, 2008, 2:58.

[5] **J. Ernst**, **G. J. Nau**, **Z. Bar-Joseph**, "Clustering short time series gene expression data," Bioinformatics, Vol. 21, No. suppl_1, 2005, pp. 159-168.

[6] **S. Ignacimuthu**, *Basic Bioinformatics*. United Kingdom:Alpha Science International Ltd., 2004, 206 p.

[7] **M. F. Ramoni**, **P. Sebastiani**, **I. S. Kohane**, *Cluster analysis of gene expression dynamics*. Proc. National Acad. Sci. USA, Vol. 99, N 14, 2002, pp. 9121-9126.

[8] **P. N. Tan**, **M. Steinbach**, **V. Kumar**, *Introduction to Data Mining*. Boston: Addison-Wesley, 2006, 769 p.

[9] **M. B. Eisen**, **P. T. Spellman**, **P. O. Brown**, **D. Botstein**, *Cluster analysis and display of genome-wide expression patterns*. Proc. National Acad. Sci., USA, Genetics, Vol. 95 (25), 1998, pp. 14863–14868.

[10] **R. R. Sokal**, **C. D Michener**, Univ. Kans. Sci. Bull. 38, 1958, pp. 1409-1438.

[11] **V. R. Iyer**, **M. B. Eisen**, **D. T. Ross**, **G. Schuler**, **T. Moore**, **J. C. F. Lee**, **J. M. Trent**, **L. M. Staudt**, **J. Hudson**, **M. S. Boguski**, **D. Lashkari**, **D. Shalon**, **D. Botstein**, **P. O. Brown**, "The transcriptional program in the response of human fibroblasts to serum", Science, Vol. 283 (1), 1999, pp. 83-87.

[12] **L. Wang**, **M. F. Ramoni**, **P. Sebastiani**, "Clustering short gene expression profiles," Lect. Notes Computer Science, 3909, 2006, pp. 60-68.

[13] **Z. Wang**, **F. Yang**, **D. V. Ho**, **S. Swift**, **A. Tucker**, **X. Liu**, "Stochastic dynamic modeling of short gene expression time-series data," IEEE transactions on nanobioscience, 7(1), 2008, pp. 44-55.

[14] **I. P. Androulakis**, **E. Yang**, **R. R. Almon**, "Analysis of time-series gene expression data: Methods, challenges, and opportunities," *Annual Review of Biomedical Engineering*, 2007, pp. 9:3.1-3.24.

[15] **A. Kirshners**, **Y. Kornienko**, "Time-series data mining for e-service application analysis," *Scientific Proceedings of Riga Technical University, Information Technology and Management Science*, Issue 5, Vol. 40, RTU, Riga, 2009, pp. 94-100.

[16] **D. C. Montgomery**, **C. L. Jennings**, **M. Kulachi**, Introduction to time series analysis and forecasting. Wiley-Interscience, 2008, 472 p.

[17] **J. J. Flores**, **R. Loaeza**, *Financial time series forecasting using a hybrid neural-evolutive approach, Proceedings of the XV SIGEF International Conference*, Lugo, Spain, 2009, pp. 547-555.

[18] **A. Kirshners**, **A. Sukov**, "Rule induction for forecasting transition points in product life cycle data," *Scientific proceedings of Riga Technical University, Information Technology and Management Science,* Issue 5, Vol.36, RTU, Riga, 2008, pp. 170-177.

[19] **S. Thomassey**, **A. Fiordaliso**, "A hybrid sales forecasting system based on clustering and decision trees," Decision Support Systems, Volume 42, Issue 1, 2006, pp. 408-421.

[20] **A. Kirshners**, **S. Parshutin**, **A. Borisov**, "Combining clustering and a decision tree classifier in a forecasting task," Automatic Control and Computer Sciences, Vol. 44, N 3, 2010, pp. 124-132.

Arnis Kirshners, Mg. sc. ing., Ph. D. Student and Senior Laboratory Assistant, Faculty of Computer Science and Information Technology, Riga Technical University, Department of Modelling and Simulation. He received his diploma from the Department of Modelling and Simulation at the Technical University of Riga. His research interests include data mining and knowledge extraction, bioinformatics, intelligent systems, programming, database and evolutionary computing. Previous publications: A. Kirshners, A. Sukov, "Rule induction for forecasting transition points in product life cycle data," *Scientific proceedings of Riga Technical University, Information Technology and Management Science,* Issue 5, Vol.36, RTU, Riga, 2008, pp. 170-177.A. Kirshners, Y. Kornienko, "Time-series data mining for e-service application analysis," *Scientific Proceedings of Riga Technical University, Information Technology and Management Science*, Issue 5, Vol. 40, RTU, Riga, 2009, pp. 94-100. A. Kirshners, S. Parshutin, A. Borisov, "Combining clustering and a decision tree classifier in a forecasting task," Automatic Control and Computer Sciences, Vol. 44, N 3, 2010, pp. 124-132. Contact information: 1 Kalku Street, Riga, LV-1658, phone: +371 67089530, e-mail: arnis.kirsners@rtu.lv.

Arkady Borisov is Professor of Computer Science in the Faculty of Computer Science and Information Technology at Riga Technical University. He holds a Doctor of Technical Sciences degree in Control in Technical Systems and the Dr. habil. sc. comp. degree. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 210 publications in the area. He has supervised a number of national research grants and participated in the European research project ECLIPS.
Contact information: 1 Kalku Street, Riga, LV-1658, phone: +371 67089530, e-mail: arkadijs.borisovs@cs.rtu.lv.

Arnis Kiršners, Arkady Borisovs. Īsu laika rindu analīze gēnu ekspresijas uzdevumos
Rakstā tiek analizētas un aplūkotas dažādas klasterizācijas pieejas, kuras tiek izmantotas gēnu ekspresijas uzdevumos. Izvēlētās pieejas ir aplūkotas un analizētas no datu ieguves klasterizācijas algoritmu izmantošanas iespēju pielietojuma. Rakstā tiek īsi aprakstīti aplūkojamo metožu un algoritmu darbības principi, raksturīgākās iezīmes un eksperimentos izmantotās datu kopas. Uzskaitīti ir eksperimentu rezultāti, kas tieši saistīti ar klasterizācijas algoritmu izmantošanu apstrādājot īsas laika rindas bioinformātikā, risinot gēnu ekspresijas uzdevumos. Izdarīti secinājumi un veikti novērtējumi par katru no izmantotajām pieejām. Analizētas nākotnes iespējas jaunas metodes izveidei, kas balstītos uz datu ieguves pieejām un principiem, bet risinātu bioinformātikas uzdevumus, kuri saistīti ar īsu laika rindu apstrādi un iegūtie rezultāti tiktu interpretēti bioinformātikas speciālistiem saprotamā veidā.

Арнис Киршнерс, Аркадий Борисов. Анализ коротких временных рядов в задачах экспрессии генов
В статье проведен анализ и обзор различных методов кластеризации, которые используются в задачах экспрессии генов. Выбранные методы анализируются и рассматриваются с точки зрения применения возможностей использования кластеризационных алгоритмов в задачах добычи данных. В статье кратко изложен принцип работы рассматриваемых методов и алгоритмов, а также - характерные черты и используемые в экспериментах выборки данных. Приведены результаты экспериментов, которые связаны с использованием кластеризационных алгоритмов, при обработке коротких временных рядов в области биоинформатики, решая задачи экспрессии генов. Сделаны выводы и произведена оценка каждого использованного подхода. Проанализированы возможные подходы к разработке в будущем нового метода, который основывается на принципах добычи данных, но в то же время решает задачи биоинформатики. Эти задачи связаны с обработкой коротких временных рядов, и полученные результаты должны интерпретироваться в виде, понятном для специалистов биоинформатики.