Доклады Международной конференции “Математическая биология и биоинформатика”. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e54. doi: 10.17537/icmbb18.81

Conceptual Approach to Clustering in the Study of Gene Expression

Bogatyrev M.Y., Samodurov K.V.

Tula State University

okkambo@mail.ru

The work contains an overview of clustering methods used in the study of gene expression and new results of application of Formal Concepts Analysis to clustering data extracted from the public functional genomics data repository GEO. Methods of Formal Concept Analysis allow clustering of multidimensional data under the single condition of partial ordering of such data sets. As a result, clusters are separate sublattices in the concept lattice, where each sublattice contains hierarchically related formal concepts. This solution of the clustering problem allows deep investigating both mutual influences of genes and their influence on other data obtained in experiments on gene expression. The paper describes a new information technology developed for the implementation of the proposed approach. The technology uses modern solutions in the field of Big Data processing, it has functions for communication with external data sources and other information systems. Preliminary results of the application of this technology to three-dimensional gene expression data obtained from the GEO system are presented.

Key words: gene expression, clustering, Formal Concept Analysis, OLAP technology.

Концептуальный подход к кластеризации в изучении экспрессии генов

Богатырёв М.Ю., Самодуров К.В.

Тульский государственный университет

Работа содержит обзор методов кластеризации, применяемых в изучении экспрессии генов, и новые результаты применения анализа формальных понятий для кластеризации данных, извлекаемых из сетевого хранилища биомедицинской информации GEO. Методы анализа формальных понятий позволяют выполнять кластеризацию многомерных данных при единственном условии частичной упорядоченности множеств таких данных. В результате кластеры представляют собой отдельные подрешётки в решётке понятий, где каждая подрешётка содержит иерархически связанные формальные понятия. Такое решение задачи кластеризации позволяет более глубоко исследовать как взаимные влияния генов, так и их влияния на другие данные, получаемые в экспериментах по изучению экспрессии генов. В работе описана новая информационная технология, разрабатываемая для реализации предложенных решений. Технология использует современные решения в области обработки больших данных, имеет функции связи с внешними источниками данных и другими информационными системами. Приведены предварительные результаты применения данной технологии к трёхмерным данным экспрессии генов, полученных из системы GEO.

Ключевые слова: *экспрессия генов, кластеризация, анализ формальных понятий, технология OLAP*.

1. Introduction

The study of gene expression has quite a long history but it is still far from being completed. The reason for this is that the problem of gene expression is fundamental one and therefore has many areas of research.

There are two aspects of gene expression problem: *technological aspect* and the aspect related to *processing of gene expression data* (*GED processing*). Technological aspect unites various techniques and tools have been used to measure gene expression in experiments. The aspect of GED processing covers wide area of research which can be treated as realization of Data Mining on gene expression data. It includes the following areas:

- gene expression data organization;
- mathematical models applied for the study of gene expression;
- methods and algorithms of GED processing.

The current trends in these three areas are the following.

Modern GED databases have VLDB (Very Large Data Base) type, they have versatile functionality and collect as raw experimental data as the results of processing them [1].

Except standard models of gene expression, some models from non-traditional for biology mathematical fields are applied: algebraic models [2], models from system theory [3], as the most striking ones.

Clustering plays significant roles in the study of gene expression. There are two crucial problems in clustering: the problem of selecting measure of similarity (proximity measure) of objects being clustered and the problem of biologically meaningful interpretation of clusters. To solve these problems, some modern approaches from data analyses have been applied: neural networks [4], Evolutionary Computations [5], bi- and triclustering [6].

This paper is devoted to GED processing by clustering. We apply Online Analytical Processing (OLAP) and triclustering technique based on Formal Concept Analysis [7].

The combination of these two techniques allows one to work with Big Data which are gene expression data.

Another important result of our method is the extension of the possibilities to interpret results of clustering, with the help of conceptual lattices.

2. Clustering in the Study of Gene Expression

In the study of gene expression, clustering techniques are applied to create groups of genes that exhibit a similar behaviour. This behavior is analyzed and evaluated by a biologist, while gene grouping is performed by a clustering algorithm which uses formally chosen proximity measure of objects being clustered. This is the main contradiction in the application of clustering to gene expression. The resolution of this contradiction generates a variety of approaches to clustering in the area of gene expression.

2.1. Gene Expression Data

Microarray technology is mostly used for obtaining GED. It has made it possible to simultaneously monitor expression levels of many thousands of genes through two major types of microarray experiments: cDNA microarray and oligonucleotide arrays [6].

Consider two variants of gene expression data organization. The first variant is real-valued gene expression matrix $\mathbf{D}_2 = \{d_{i,j}\}$.The rows of this matrix form the expression patterns of genes, the columns represent the expression profiles of samples, and each cell $d_{i,j}$ is the measured expression level of gene i in sample j.

The second variant of gene expression data organization is based on the known results of the interpretation of gene expression. Through gene expression, genes are linked to each other in a particular organism, or they linked with certain diseases, or another target of linking may be selected. The data of such kind may is obtained as a result of query to GED databases [1]. Depending on the number of queried parameters, the output data may have more general form of tensor: $\mathbf{D}_n = \{d_{i,j,\ldots,z}\}$.

Tensors of no higher than three dimensions are used in the gene expression analysis. This is due to the significantly increasing computational difficulties in processing such data.

There are two variants of clustering GED on the matrices of $\mathbf{D}_2$ type: ordinary clustering and biclustering. Triclustering is applied for GED which has the form of three-dimensional tensor $\mathbf{D}_3 = \{d_{i,j,k}\}$.

2.2. Bi- and triclustering of GED

Ordinary clustering [6] was the main variant of GED clustering during the years. It is based on the clustering the single set of data. In the *gene-based clustering*, this set contains genes in the $\mathbf{D}_2$ matrix and they are treated as the objects, while the samples in that matrix are the features or parameters of objects. Another variant is the *sample-based clustering* when the samples are the objects and the genes are the features. Both these approaches face the following challenge.

Gene expression data often contain a great amount of noise. This and other features of the data cause clusters to actually intersect.

Bi- and triclustering represent possible solutions to this problem. These two approaches are based on the following idea. If objects and features are linked then clustering is performed on the both these sets. Triclustering expands this proposition to three sets. In other words bi- and triclustering are two- and three-dimensional clustering respectively.

First biclustering algorithm was proposed in [8]. Biclusters are sub matrices in the original matrix of GED which meet the demands to have so called *the mean-squared residue* less or equal to certain value [8]. That characteristic of clusters serves as proximity measure.

For a gene expression matrix containing n genes and m samples, the computational complexity of a complete combination of genes and samples is 2^{n+m} so that the problem of optimal block selection in the gene expression matrix is *NP*-hard [6]. This is actual for biclustering.

Triclustering is an expansion of biclustering which uses different time points as third dimension together with dimensions of genes and experimental conditions. Triclustering can be applied to 3D Microarray Data [9]. As for biclustering, the problem of computational complexity dramatically increases for triclustering. This is the main reason for finding new triclustering algorithms that reduce the severity of this problem.

3. Conceptual Approach to GED Clustering

Conceptual approach based on Formal Concept Analysis is an alternate method of clustering which is applied as for bi- as for triclustering [10].

3.1. Formal Concept Analysis

Formal Concept Analysis (FCA) [7] is the paradigm of conceptual modeling which studies how objects can be hierarchically grouped together according to their common attributes. That grouping of objects is really biclustering of them. The output of FCA algorithms is *conceptual lattice* which contains hierarchically linked *formal concepts* which are biclusters.

Formal Concept Analysis has been constructively applied for knowledge discovery from biological data [11–13].

Classical FCA uses so called *formal context* which is given by a binary matrix. *Formal triadic context* and subsequent notions of FCA are determined as following.

Formal triadic context is quadruple $\mathbf{K} = (G, M, B, I)$ [14] which elements are: G is a set of objects, M – set of their attributes, B – set of conditions under which G has attributes M, $I \subseteq G \times M \times B$ is a triadic incidence relation. All the sets of G, M, B are partially ordered.

A triadic concept is a triple (A_1, A_2, A_3) such that $A_1 \subseteq G, A_2 \subseteq M, A_3 \subseteq B$ and for every $\{i, j, k\} = \{1, 2, 3\}$ with $j < k$ $(A_j \times A_k)^{(i)} = A_i$. Here $(.)^{(i)}$ is the *derivation operator* which maps appropriate subsets from the sets G, M, B into one another [15]. It is true that formal triadic context is three-dimensional [0, 1]-tensor, which under suitable permutations of rows, columns, and layers may be transformed so that every triadic concept (A_1, A_2, A_3) is interpreted as a maximal cuboid full of units. Every triadic concept is tricluster but the opposite is not true [15].

There are several approaches to constructing triclustering FCA algorithms and studies relations between triadic concepts and triclusters [15].

3.2. FCA – OLAP Technology

We are developing the new technology of triclustering which has the following characteristics:

1. it communicates online with public GED databases including Gene Expression Omnibus [1];
2. it uses OLAP technology [16] for processing data;
3. it is based on FCA triclustering algorithms and uses triadic concepts for interpreting results of triclustering.

Communicating with Gene Expression Omnibus we can get gene expression data as of the first as of the second variants – see paragraph 2.1.

OLAP is the set of tools intended to analyze multidimensional data interactively from multiple perspectives. It uses multidimensional cube for data presentation and processing. Such cube is generated from the data stored in a database or from external data source.

Developing technology interacts with FCART [17], the modern system realizing some FCA algorithms.

3.3. Experiments and current results.

The OLAP itself is not fast for data processing. So we needed to realize it under modern database technology oriented on Big Data. We use SAP-Sybase IQ database server [18] which satisfies the requirements for processing Big Data and supports OLAP operators in its SQL.

We realized modified version of *OAC-triclustering* algorithm [19] which has linear time and memory complexities and is effective for triclustering Big Data.

Among our experiments with GED from GEO [1] the mostly interesting are those in which we studied cross interactions between genes, organisms and diseases. To realize such study, special queries to GEO were constructed and special tools for filtering and processing output data were programmed too. A characteristic feature of the output data is their non-numeric nature and the presence of whole phrases in the output set. The following fragment from output data set illustrates that feature:

< *ampylobacter, Cj1456c, Flagellar export apparatus component FlhA inactivation effect on virulence*>.

It has the template: <*organism, gene, information about disease* >.

The following table illustrates summary results of the study of the interaction of genes, organisms and diseases.

Table 1. Summary results of genes interactions

Organisms	Number of concepts	Number of genes	Number of signs of diseases
Bacteria (720)	76893	2832640	432
Mammals (346)	13264	4265482	578
Marsupials (84)	1287	92678	257

We restricted the number of analyzed organisms to the values shown in the first column of the table. In the rows of the table there are number of concepts in which certain numbers of different organisms are presented and corresponded them number of genes and number of signs of diseases are presented too. Here we have three sub lattices “*Bacteria*”, “*Mammals*” and “*Marsupials*” which are linked but not intersect.

We had similar result presenting clusters as disjoined lattices in the study of Gram properties of bacteria.There we used visualization of lattices by special *views* [13]. In the current research visualization is not helpful due to the giant size of conceptual lattice.

4. Conclusion an further research

This work is devoted to developing clustering technique for the study of gene expression. We apply FCA as an alternate method of clustering which uses partial ordering instead of numerical proximity measure. This more general kind of proximity measure allows obtaining more general results of gene expression. Using concept lattices it is possible to study cross interactions between genes and other essences, for example organisms and diseases.

Further steps in the developing proposed technology are the following.

1. Creating interactive interface for interpretation formal concepts in the lattice. Currently visualization is the main instrument for interpretation formal concepts in the conceptual lattice [5–7]. But if a lattice is big as in our experiments then visualization is not effective. It should be supplemented by special domain-oriented interface. This interface has special tools which help a biologist to classify and to interpret data which constitute formal concepts.

2. The OLAP does not restrict dimension of the data cube. So it is possible to realize *n*-clustering approaches based on *n*-adic formal context and corresponding multidimensional conceptual lattices [20]. That modification of the technology will provide deeper analysis of the gene expression data.

5. References

1. *GEO: Gene Expression Omnibus. Public functional genomics data repository*. URL: https://www.ncbi.nlm.nih.gov/geo/ (accessed 16.08.2018).
2. Aleem H.A., Mavituna F., Green D.H. A Galois Field Approach to Modelling Gene Expression Regulation. In: *Proceedings 38th International Symposium on Multiple Valued Logic*, 2008. P. 88–93.
3. Dini P., Egri-Nagy A., Nehaniv C., Schilstra M. *Mathematical Models of Gene Expression Computing. OPAALS Project Report.* 77 p. URL: http://www.lse.ac.uk (accessed 16.08.2018).
4. Tan A.H., Pan H. Predictive neural networks for gene expression data analysis. *Neural Netw*. 2005. V. 18. № 3. P. 297–306.
5. Ma P.C.H., Chan K.C.C., Yao X., Chiu D.K.Y. An Evolutionary Clustering Algorithm for Gene Expression Microarray Data Analysis. *IEEE Transactions on Evolutionary Computation*. 2006. V. 10. № 3. P. 296–314.
6. Jiang D., Tang C., Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*. 2004. V. 16. № 11. P. 1370–1386.
7. Ganter B., Stumme G., Wille R. *Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2005. № 3626.
8. Cheng Y., Church G.M. Biclustering of expression data. In: *Intelligent Systems for Molecular Biology (ISMB):* Proceedings of the Eighth International Conference. 2000. V. 8. P. 93–103.
9. Zhao L., Zaki M.J. Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. In: *Management of data:* Proceedings of the 2005 ACM SIGMOD international conference. 2005. P. 694–705.
10. Kaytoue M., Kuznetsov S.O., Napoli A., Duplessis S. Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*. 2011. V. 181. № 10. P. 1989–2001.
11. Raza K. Formal Concept Analysis for Knowledge Discovery from Biological Data. *International Journal of Data Mining and Bioinformatics*. 2017. V. 18. № 4. P. 281–300.
12. Bogatyrev M.Yu., Vakurin V.S. Conceptual Modelling in the Study of Biomedical data. *Mathematical Biology and Bioinformatics*. 2013. V. 8. № 1. P. 340–349 (in Russ.).
13. Bogatyrev M.Y. On the Study of Bacteria Biotopes with Formal Concept Analysis. In: *Mathematical Biology and Bioinformatics:* Proc. International conference. Pushchino: MAX Press, 2016. C. 120–121.
14. Lehmann F., Wille R. A triadic approach to Formal Concept Analysis. In: *Conceptual structures: Applications implementation and theory:* Proceedings of the third international conference. London: Springer, 1995. P. 32–43.
15. Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. In: *Machine Learning*. 2015. P. 1–32.
16. Thomsen E. *OLAP Solutions: Building Multidimensional Information Systems*. Willey &Sons, 2002. 661 p.
17. Neznanov A., Kuznetsov S.O. Information Retrieval and Knowledge Discovery with FCART. In: *Formal Concept Analysis Meets Information Retrieval:* Proceedings of the Workshop. 2013. P. 74–82.
18. *Columnar Analytics Database Software*. URL: https://www.sap.com/products/sybase-iq-big-data-management.html (accessed 16.08.2018).
19. Gnatyshak D., Ignatov D.I., Kuznetsov S.O., Nourine L. A One-pass Triclustering Approach: Is There any Room for Big Data? In: *Concept Lattices and Their Applications (CLA 2014):* Proc. 11th International Conference. Eds. Bertet K., Rudolph S. 2014. V. 1252. P. 231–242.
20. Voutsadakis G. Polyadic concept analysis. *Order*. 2002. V. 19. № 3. P. 295–304.