Clustering of Multidimensional Objects in the Formation of Personalized Diets

Valentina N. Ivanova[1]
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Igor A. Nikitin[2]
Department "Technology of grain processing, bakery, pasta and confectionery industries"
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Natalia A. Zhuchenko[3]
Department "Medical Genetics"
I.M. Sechenov First Moscow State Medical University (Sechenov University)
Moscow, Russian Federation

Marina A. Nikitina[4]
V.M. Gorbatov Federal Research Center for Food Systems
Moscow, Russian Federation

Yury I. Sidorenko[5]
Department "Technologies of production and organization of catering and merchandising
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Vladimir I. Karpov[6]
Department "Information systems and technologies"
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

Igor V. Zavalishin[7]
K.G. Razumovsky Moscow State University of technologies and management (The First Cossack University)
Moscow, Russian Federation

***Abstract*—When developing personalized diets (personalized nutrition) it is necessary to take into account individual physiological nutritional needs of the body associated with the presence of gene polymorphism among consumers. This greatly complicates the development of rations and increases their cost. A methodology for the formation of target diets based on the multidimensional objects clustering method has been proposed. Clustering in the experimental group was carried out on the basis of a calculation of the integral assessment of reliable risks of developing decease conditions according to selected metabolic processes. And genetic data of participants was taken into account. The use of the proposed method allowed reducing the needed number of typical solutions of individual diets for the experimental group from 10 to 3.**

Keywords—Multidimensional objects clustering method; integral assessment of reliable risks; nutritional needs of the body; personalized nutrition

I. Introduction

Modern studies of the human genome have allowed the identification of many genes responsible for metabolic processes, whose polymorphism plays a significant role in the occurrence of metabolic disorders and the development of diseases. Identifying the alleles of such genes that are present in humans helps to determine the risk factors of particular health disorders and to develop optimal measures that will prevent the negative influence of environmental factors on the implementation of genetically determined disorders [1].

One of decisive factors determining the diet is the human genome [2]. Today, a reliable statistical relationship has been established between the presence of certain varieties (alleles) of fixed genes in relation to susceptibility to more than 150 hereditary diseases [3]. The process of occurrence of a disease may be associated with disruptions in the functioning of individual organs and systems and be a consequence of a violation of nutritional status, which does not take into account the peculiarities of the genetic influence on the nutrient needs of the body. Thus, food products and food rations, designed to meet the corrected needs for food nutrients that take into account genetic characteristics of a particular organism, automatically prevent the adverse functioning of problem organs and systems [4-6].

The use of statistical methods for analyzing medical information is currently relevant. With the development of technology, the sphere of their use is expanding and includes the methods of information processing called Data Mining.

One of the main effective and widely used methods of Data Mining in relation to large amounts of information is a clustering method. The point of the method is in searching signs of similarity between objects in a particular subject area and the subsequent merging of objects into subsets (clusters) according to established signs of similarity.

Data mining contains methods of detection, data collection, as well as its intellectual analysis. Data Mining is a multidisciplinary field that emerged and develops on the basis of such sciences as applied statistics, pattern recognition, artificial intelligence, database theory, etc.

This study examined the effect of a limited list of gene panels on metabolic processes with the calculation of the integral assessment of reliable risks for the development of disease conditions, and also proposed the application of the multidimensional objects clustering method in order to form diets for target groups of consumers.

The task of clustering is due to the fact that in the case of mass (industrial) formation of rations, the problem of finding typical solutions arises. These solutions should be made for target groups of consumers assigned to a particular cluster. It should be noted that clusters themselves are unknown in advance. Therefore, in order to accumulate information about clusters during scientific research, the clustering problem is solved and the method of their formation is worked out [7, 8].

II. Research Methodology

The group included people of European type (men and women), about the same age (28-35 years old), born and living in several generations in the region of Central Russian upland. The polymorphisms of genes involved in the main metabolic processes and causing the risk of occurrence of certain diseases were selected as the most significant ones: biotransformation of xenobiotics, metabolism of vitamins, assessment of psycho-emotional status.

Table 1 lists the controlled alleles of genes and corresponding risks of hereditary multifactorial diseases for the mentioned above metabolic processes.

Biotransformation of xenobiotics is a biochemical process during which substances transform under the action of various enzymes of the body [9-17]. Its biological meaning is the transformation of a chemical substance into a form suitable for removal from the body. Four genes of the activation phase of xenobiotics (CYP1A1*2B, *4, CYP2D6*3, *4, CYP2C9*2, *3 and CYP2C19*2) and four genes of the detoxification phase (GSTT1, GSTM1, NAT2 and TPMT) were included in the biotransformation panel under study. To assess the vitamin status of the organism, marker genes that indicate risks of reducing the concentration of vitamins in the organism of the genome carrier (NBPF3 (ALPL), FUT2, BCMO1, APOA5) were studied [18, 19]. To assess the psychoemotional status of participants in the experimental group, gene activities (DRD-2A, SR (HTR2A)) responsible for the synthesis of serotonin and dopamine enzymes were also identified [20, 21].

These gene panels are associated with a predisposition to a number of most common diseases and are included in the list of genetic tests of most medico-genetic laboratories.

TABLE I. The List of Reliable Risks of Pathologies Corresponding to Selected Metabolic Processes

Metabolic process encoded by a group of genes		**The name of the polymorphism gene carrier**	**Risk of pathology / disease**
Biotransformation of xenobiotics	Phase 1 - activation	CYP1A1*2B,*4	Lung cancer, acute leukemia, general oncology, proton pump inhibiting
		CYP2D6*3,*4	Metabolism of psychotropic drugs, including drugs of a narcotic series
		CYP2C9*2, *3	Metabolism of antidepressants, β-adrenoreceptor blockers
		CYP2C19*2	Metabolism of some pharmaceuticals, including proton pump inhibitors
	Phase 2 - Detoxification	GSTT1	Bowel Cancer. Encode the synthesis of the enzyme glutathione-S-transferase. Activate glutathione
		GSTM1	Bowel Cancer. Encode the synthesis of glutathione-S-transferase. Activate glutathione
		NAT2	Encodes the enzymes responsible for the catalysis of aromatic xenobiotics by acetylation. Determines the rate of occurrence of a malignant neoplasm of the walls of the bladder and rectum
		TPMT	Responsible for the synthesis of the enzyme thiopurine-S-methyltransferase, which is associated with the processes of detoxification of the body.
Vitamin metabolism		NBPF3(ALPL)	The risk of reducing the concentration of vitamin B6
		FUT2	The risk of reducing the absorption of vitamin B12
		BCMO1	Risk of disorders in vitamin A synthesis from β-carotene
		APOA5	Risk of low levels of α-tocopherol (vitamin E)
Psycho-emotional status		DRD-2A	The formation of addiction to alcohol and narcotic substances due to a deficiency in the synthesis of serotonin and dopamine.
		SR(HTR2A)	Associated with increased risk of paranoid schizophrenia

Experiment participants were assigned reference numbers from 1 to 10. Testing was performed by analyzing saliva using the micronucleus test of the buccal epithelium. As a result of testing, data on the presence of polymorphisms in the homozygous safe (C / C), heterozygous (C / A) or homozygous predisposing to the disease (A / A) forms in the studied genes was obtained. For the ease of processing the experimental data the presence of polymorphism in the homozygous form predisposing to the disease was indicated by a score of 2 points, in the heterozygous form by a score of 1 point, and the homozygous safe form by a score of 0 points. Table 2 shows information on the presence of polymorphisms in genes tested in the experiment or their alleles in one form or another.

Table 2 shows the individual and integral assessment of reliable risks of expression of genes and their alleles tested in the experiment. This table is compiled in the form of a matrix. The sum of points accumulated by each participant on the studied gene alleles expressed an integral risk assessment for each participant in the experimental group (ranging from 0 to 30).

Summary line of the sum of risks for each group member given in Table 2 allows to give an integrated risk assessment of diseases of the whole spectrum of diseases determined by considered gene panels.

Mathematical data processing was performed using soft calculations, namely clustering of multidimensional objects [22-27].

TABLE II. ESTIMATION OF RELIABLE RISKS OF THE PROBABILITY OF DEVELOPING DECEASE CONDITIONS BY SELECTED METABOLIC PROCESSES, EXPRESSED IN POINTS (HIGH PROBABILITY–2 POINTS, MEDIUM – 1 POINT, LOW – 0 POINTS)

Metabolic process encoded by a group of genes		**The name of gene**	**Gene sequence number**	**Number in the group**									
				1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
				Sex									
				m	***m***	***m***	***m***	***m***	***m***	***m***	***m***	***m***	***f***
Biotransformation of xenobiotics	Phase 1 - activa-tion	CYP1A1*2B,*4	1	0	1	1	0	1	1	0	0	0	1
		CYP2D6*3,*4	2	0	0	0	1	0	0	1	0	0	0
		CYP2C9*2, *3	3	0	1	1	0	0	1	0	0	0	0
		CYP2C19*2	4	1	0	0	0	0	0	1	0	0	1
	Phase 2 – Detoxi-fication	GSTT1	5	2	0	2	0	0	2	0	0	2	0
		GSTM1	6	2	0	0	0	2	0	0	2	0	2
		NAT2	7	2	1	2	2	1	1	1	1	2	2
		TPMT	8	0	0	0	0	0	0	0	0	0	0
Vitamin metabolism		NBPF3(ALPL)	9	1	1	2	1	1	1	1	2	1	2
		FUT2	10	1	2	2	2	2	1	0	1	2	2
		BCMO1	11	2	2	0	0	1	0	1	1	0	0
		APOA5	12	2	2	2	2	2	2	2	2	2	2
Psycho-emotional status		DRD-2A	13	0	0	0	0	0	0	0	0	0	0
		SR(HTR2A)	14	2	1	1	2	2	2	2	1	2	1
Integral evaluation				16	11	14	11	13	12	11	11	12	13

III. Mathematical Formulation of the Clustering Problem

Given:

C_0 - the initial set of objects of study,

C0= {Sn}, n = 1, ..N

Mp(M) – metrics of characteristics

Mp(i)–the weight of importance of risk at the i-th gene condition, 1,.. M

X(n, i)–risk assessment in points in accordance with condition of the i-th gene in object n, n=1, ..N, i =1,..M, $\forall n \forall i$ X(n, i) $\in$ {0,1,2}

The metric Mp (M) is normalized.

$$\sum_{i=1}^{N} Mp(i) = 1 \tag{1}$$

The initial set C_0 must be divided into sets of clusters C_k:

C_0={ C_k } k=1,..K (2)

C_k = { S_z }, z =1,..N_k (3)

Any pair of clusters has no common elements, that is, any object can only be in one cluster;

$$\forall C_k \in C_0, \forall C_l \in C_0 : C_k \bigcap C_l = \varnothing \tag{4}$$

It is required to determine such C_k that maximize the criterion *U*:

$$U(\mathrm{K}_o) = \max_{K=\overline{N,2}} \{\mathrm{U}_1(K) - U_2(K)\} \tag{5}$$

Where *U(K$_o$)* is the optimal value of the clustering quality criterion;

U$_1$(K) - compactness of classes with *K* clusters;

U$_2$(K) is a measure of similarity of classes with *K* clusters.

The measure of similarity between two objects is determined on the basis of the potential function *f (S$_i$, S$_j$):*

$$f(S_i, S_j) = \frac{1}{1+\rho^2(S_i, S_j)} \ ,$$

$$\rho(S_i, S_j) = \sqrt{\sum_{m=1}^{M} (\mathrm{Mp(m)} * (X_{im} - X_{jm}))^2}$$

$$U_1(K) = \frac{1}{K}\sum_{k=1}^{K}\frac{2}{N_k(N_k-1)}\sum_{S_i\in C_k}\sum_{S_j\in C_k} f(S_i,S_j)\,, \quad i \neq j$$

where *K* is the number of classes at the current classification step;

C$_k$ – k -th class of objects;

N$_k$ - the number of objects in the class *C$_k$*;

f (S$_i$, S$_j$) - potential function of two objects *S$_i$* and *S$_j$;*

(S$_i$, S$_j$) - the distance between objects *S$_i$* and *S$_j$* in the space of characteristics *X*, taking the metric into account

$$F(C_k, C_l) = \frac{1}{N_k N_l}\sum_{S_i\in C_k}\sum_{S_j\in C_l} f(S_i,S_j)$$

$$U_2(K) = \frac{2}{K(K-1)}\sum_{C_k\in \boldsymbol{C_p}}\sum_{C_l\in \boldsymbol{C_p}} F(C_k, C_l)\,, \quad k \neq l$$

Thus, optimal splitting into clusters implies maximizing the criterion *U(K$_o$)* (see formula 5). Substantially such a statement means that in each cluster related objects are collected, and between objects of different clusters there are significant differences. This problem is related to soft computing problems class solved by the methods of integer mathematical programming. To solve the problem a set of programs for assessing the quality of multidimensional objects was used [9].

IV. Results and Discussion

When solving the clustering problem on the example of the study group, four metabolic processes were distinguished:

biotransformation of xenobiotics-activation phase (process number 1);

biotransformation of xenobiotics-detoxification phase (process number 2);

metabolism of vitamins (process number 3);

assessment of psychoemotional status (process number 4).

Each process is encoded by several genes (from two genes in the psycho-emotional status, up to four in each of other processes). A possible condition for the clustering of participants is the presence of approximately the same total number of points within each process and, accordingly, close values of integral assessments of reliable risks for the amplification of disease states on selected metabolic processes.

Table 3 provides information on the integral assessment of reliable risks for the above mentioned processes:

process number 1: genes numbered 1, 2, 3, 4;

process number 2: genes with numbers 5, 6, 7, 8;

process number 3: genes numbered 9, 10, 11, 12;

process number 4: genes numbered 13, 14.

TABLE III. Integral Assessment of Significant Risks for Selected Metabolic Processes

The process number encoded by the gene group	**Member number in the group**									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
	Integral assessment of reliable risks for selected processes for each participant									
Process 1	1	2	2	1	1	2	2	0	0	2
Process 2	6	1	4	2	3	3	1	3	4	4
Process 3	6	7	6	5	6	4	4	6	5	6
Process 4	2	1	1	2	2	2	2	1	2	1

As a result for each participant in the experiment, the sum of the risks for each of the four processes is calculated. For example, for the first participant we get an integral assessment of reliable risks for process # 1: 0 + 0 + 0 + 1 = 1, for process # 2: 2 + 2 + 2 + 0 = 6, for process # 3: 1 + 1 + 2 + 2 = 6, etc.

The problem of combining objects into clusters based on the data from Table 3 was solved according to the condition that integral assessments of reliable risks in processes differ by no more than 25% among the participants of one cluster.

The results of solving the clustering problem are given in Table 4.

Table 4 shows that the number of individual decisions for which specialized menus should be made reduced from 10 to 3. That is participants numbered 9, 4, 2, 7 and 5 are assigned to cluster 2 (integral risk is in the range of 0.60 to 0.71), participants 3, 10, 6 and 1 are assigned to cluster 3 (integral risk is in the range of 0.76 to 0.84). Participant 8 is assigned to an independent cluster 1 (integral risk–0.46).

Table 4 also provides information on the integral risk in the form of a conditional value from 0 to 1 for each member of the group, where zero corresponds to the presence of polymorphisms in the homozygous safe form in all 14 genes in the alleles under study, and 1 corresponds to the presence of polymorphisms in the homozygous form that predisposes a disease in each of the 14 genes.

Using intelligent data processing with clustering methods, you can simulate a personalized optimal diet for a participant based on medical indicators in terms of minimizing the risk function. As can be seen in Table 4 NAT2 and APOA5 genes make the greatest contribution to the risks of hereditary diseases for people assigned to cluster 3. Therefore, the cluster 3 consumer group nutrition ration must necessarily take into account the corrected nutritional requirements associated with these genes.

TABLE IV. THE RESULT OF COMBINING OBJECTS (PARTICIPANTS) INTO CLUSTERS

Item number	**Participant number in group**	**Cluster 1 Integral risk**	**Cluster 2 Integral risk**	**Cluster 3 Integral risk**
1	Participant 8	0,46	-	-
2	Participant 9	-	0,60	-
3	Participant 4	-	0,64	-
4	Participant 2	-	0,67	-
5	Participant 7	-	0,68	-
6	Participant 5	-	0,71	-
7	Participant 3	-	-	0,76
8	Participant 10	-	-	0,76
9	Participant 6	-	-	0,77
10	Participant 1	-	-	0,84

The NAT2 gene is responsible for the detoxification of xenobiotics. It reduces the enzymatic activity of a number of enzymes and provokes colon and bladder cancer. In this regard, the diet of participants in cluster No. 3 should additionally contain food enriched with natural and engineered antioxidants. Since this gene also plays an important role in the detoxification of pesticides and in carcinogenesis processes, people with a high risk for this gene should prefer organic food and be attentive to products that can accumulate pesticides and heavy metals.

The APOA5 gene regulates the level of α-tocopherol (vitamin E). For people with an unfavorable genotype for this gene, it is necessary to increase the intake of vitamin E by eating more foods with a high content of it.

In cluster 2, the most provocative genes are APOA5 and SR (HTR2A). The SR gene (HTR2A) encodes the synthesis of serotonin, affecting the psychological stability of the consumer. It is possible to increase the level of serotonin by enriching the diet with offal, group B vitamins, Ca and Mg macronutrients.

In cluster 1 genes GSTM1, NBPF3 and APOA5 make the greatest contribution to the risks of hereditary diseases. Cluster number 1 participant is recommended to eat foods high in vitamin E, wholemeal bread, bran and nuts.

V. CONCLUSION

On an example of the genome analysis of the considered consumer group, a methodology was developed for the formation of target diets based on multidimensional objects clustering method. Using Data Mining (clustering method) allows to construct a balanced daily ration for personalized nutrition. Based on the study, data collection, compilation and processing of numerical information based on medical indicators, it reduces the number of rations being developed from 10 to 3.

On the base of genetic data of experimental group participants included in one or another cluster, the development of the diet of the target group should take into account adjusted physiological needs for food nutrients associated with the presence of gene polymorphism of these participants.

ACKNOWLEDGMENT

The authors are grateful to the staff of the Center for Personalized Medicine of the Sechenov First Moscow State Medical University for assistance in conducting experiments related to the definition of polymorphisms indicated in the article, as well as to students of "Molecular Dietology" specialty for participating in an experiment.

REFERENCES

[1] 1000 Genomes Project Consortium A, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al.: A global reference for human genetic variation. Nature 2015, 526:68-74.

[2] Stavros Bashiardes, Anastasia Godneva, Eran Elinav, and Eran Segaward. Towards utilization of the human genome. Current Opinion in Biotechnology 2018, 51: 57–63.

[3] Fallaize R., Macready A.L., Butler L.T., Ellis J.A., Lovegrove J.A.: An insight into food acceptance nutrient-based personalized nutrition. Nutr Res Rev 2013, 26: 39-48.

[4] Shenderov, B.A. “Omik” - technologies and their significance in modern preventive and restorative medicine / B.A. Shenderov // Bulletin of restorative medicine. - 2012.-№3, p.70-76.

[5] Ivanova, V.N. Development of the methodology for forming diets for target groups of consumers based on the analysis of their genomes / V.N. Ivanova, N.A. Zhuchenko, I.A. Nikitin, M.Yu. Sidorenko, S.V. Shterman, A.Yu. Sidorenko // Food industry.-2018. - № 10. - p. 40 - 44.

[6] Baturin A.K., Sorokina E.Yu., Pogozheva A.V., Tutelyan V.A. Genetic approaches to personalization of food // Nutrition issues. 2012. V. 81, No. 6. P. 4-11.

[7] Kulinsky, V.I. Neutralization of xenobiotics / V.I. Kulinsky // Soros Educational Journal. 1999 - №1. - P.8-12.

[8] V.N. Ivanova, V.I. Karpov, Yu.I. Sidorenko, N.A. Zhuchenko. The problem of genotypes clustering in the decision-making support system in the management of personalized nutrition // Reports of the XXI International Conference on Soft Computing and Measurements (SCM-2018). St. Petersburg. May 23-25, 2018 SPb. St. Petersburg Electrotechnical University "LETI". volume 2, section 7. Applications of decision support systems in the economy and the social sphere, pp. 303-307.

[9] Comprehensive quality assessment and classification of multidimensional objects. // Certificate of the Russian Federation on computer program official register, number 2006613936; Myshenkov K.S., Karpov V.I., Getman V.V. - No. 2006613704; Requested November 2, 2006; Registered November 16, 2006.

[10] Polonikov, A.V. Ecological and toxicogenetic concept of multifactorial diseases: from understanding etiology to clinical use / A.V. Polonikov, V.P. Ivanov, M.A. Solodilova // Medical genetics: a monthly scientific journal. - 2008 - Volume 7, N 11. - p. 3-20.

[11] Simon, V.A. Cytochrome P450 and interaction of medicinal substances / V.A. Simon // Russian Journal of Gastroenterology, Hepatology, Coloproctology .- 2002 -№6. - C.25-9.

[12] Ahsan, H. Ahsan, A.G. Measuring of the genotype versus gene products. Rundle // Carcinogenesis. - 2003 - Vol. 24, №9. - P. 1429-1434.

[13] Saprin, A.N. Metabolism and detoxification enzymes of xenobiotics / A.N. Saprin // Advances in biological chemistry.-1991-T.32-p. 146-172.

[14] Polonikov, A.V. Genetic variation of genes for xenobiotic-metabolizing enzymes and risk of bronchial asthma: The importance of gene-gene and gene-environment interactions for disease susceptibility / A.V.Polonikov, V.P. Ivanov, M.A. Solodilova // Journal of Human Genetics - 2009 - 54 (8). - P.440-449.

[15] Gilliland, F.D. Effect of glutathione S-transferase Ml and PI genotypes on xenobiotic enhancement of allergic responses: randomized, placebo-controlled crossover study / F.D. Gilliland, Y. F. Li., A. Saxon, D. Diaz-Sanchez // Lancet. - 2004 - Vol.363. - P.119-125.

[16] Hayes, J.D. Glutathione transferases / J.D. Hayes, J.U. Flanagan, I.R. Jowsey // Annu. Rev. Pharmacol. Toxicol. - 2005 - Vol.45. - P.51-88.

[17] Khudoley, V.V. Carcinogens: characteristics, patterns, mechanisms of action. SPb: Research Institute of Chemical Technology and University. - 1999 - 419 p.

[18] Egorenkova N.P., Pogozheva A.V., Sorokina E.Yu., Peskova E.V. et al. Study of metabolic peculiarities in individuals with rs9939609 polymorphism of the FTO gene // Nutrition Issues. 2015. V. 84, No. 4. P. 97-104.

[19] EFSA NDA Panel (EFSA Panel on Dietetic Products Nutrition and Allergies). Scientific opinion on dietary reference values for folate. EFSA J. 2015, 12, 3893.

[20] Leonard B. Mechanical Mechanism, Physical Remediation and Oxidation and Nitrosative Depression / B. Leonard, M. Maes // Neurosci Biobehav Rev. - 2012. - V. 36 - № 2. - P. 764-785.

[21] Alfimova M.V. Gene polymorphism of the serotonin receptor (5-HTR2A) idisbindin (DTNBP1) and the individual components of the short-term hearing-speech memory in schizophrenia / M.V. Alfimova, M.V. Monakhov, L.I. Abramova, S.A. Golubev, V.E. Golimbet // Journal of Neurology and Psychiatry. - 2009, p.70-75.

[22] The micronucleus test of the buccal epithelium of the human oral cavity: problems, achievements, prospects / V.N. Kalaev, V.G. Artyukhov, M.S. Nechaev // Cytology and Genetics. - 2014. - Vol. 48, No. 6. - P. 62-80.

[23] Estivill-Castro, Vladimir (20 June 2002). "Why so many clustering algorithms – A Position Paper". ACM SIGKDD Explorations Newsletter. 4 (1): 65–75. doi:10.1145/568574.568575.

[24] Kaufman L., Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. – Hoboken, New Jersey: John Wiley & Sons, Inc., 2005. – 355 p.

[25] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[26] Sugar C., James G. Finding the number of clusters in a data set: An information theoretic approach // Journal of the American Statistical Association. – 2003. – № 98(463). – pp. 750–763.

[27] Ben-Hur A., Guyon I. (2003) Detecting Stable Clusters Using Principal Component Analysis. // In: Brownstein M.J., Khodursky A.B. (eds) Functional Genomics. Methods in Molecular Biology. 2003. – Vol 224, Humana Press. – pp. 159-182. https://doi.org/10.1385/1-59259-364-X:159.