UDC 004:591.5:612:616-006 https://doi.org/10.15407/biotech10.05.005

CLUSTER ANALYSIS IN BIOTECHNOLOGY

O. M. KLYUCHKO

Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology of the National Academy of Sciences of Ukraine, Kyiv

E-mail: kelenaXX@ukr.net

Received 27.06.2017

The goal of publication was the analysis of cluster methods and possibility of their application in biotechnology. The evidences found in scientific literature were summarized and analyzed. This article gives a brief description of cluster analysis — basic principles, some examples of their application are given for biotechnological problems. Results of the biotechnological studies that required application of cluster methods in combination with other mathematical approaches are considered. The conclusion contains an evaluation of the performed analysis as well as recommendations on the application of cluster analysis methods in biotechnology.

Key words: cluster analysis, biotechnology.

Cluster analysis began to be used in biology since the end of the XX century. Although at the beginning they were only narrowly applied and allowed solving only a few rather specific tasks, modern clustering methods successfully deal with a lot of biotechnological problems [1], such as separating similar-looking objects, their classification, selecting some objects among others, etc. Complications arise if it is to be done reliably and mathematically reasonably, or to further software development [1–4].

Such methods are useful, for example, if we would like to separate discrepant cells *in vitro*, when the changes are still practically invisible (for example, to distinguish living and abnormal cells — dead, malformed, and those that ditter in any other way, etc.). Another example is reliable computer separation of some types of macromolecules with slightly changed chemical structures. Contemporary biotechnology faces difficulties in solution of hundreds of such tasks. Currently, it requires new approaches or modifications of previously used methods [1–4]. A novel and very attractive idea is to use cluster methods to solve modern problems, for example, to develop databases in biology and medicine; when into different fields of tables it is necessary to include the objects with only little differences from each other [1]. This field needs tight collaboration between biologists and mathematicians, who have invented numerical approaches that could be useful for biotechnological tasks; sometimes it is very difficult to pick the suitable method among them.

There is no sense in briefly describing and evaluating all types of cluster analysis methods given in so many articles (this is a task for thick volumes). So, the author decided to analyze all existing cluster methods with the aim to find regularities and trends for their use in biotechnology. Fortunatelly, the high level of math abstractions permits to reveal major tendencies. We see such abstraction with further cluster methods' classification as good help for biotechnologists for solution of their individual work tasks. We provide (1) examples of different tasks in biotechnology solved using these methods, (2) a thorough large-scale review of contemporary clustering methods applicable to biotechnology, and (3) modern classifications of clustering methods that differ in some groups of countries. Further we suggest (4) an example of task solution of practical value using the methods of cluster analysis on cells' differences. We also evaluate clustering analysis' applicability for some problems. It should make it easy to use the methods in biotechnology.

Evaluability of technical and biotechnical systems as applications of cluster analysis methods (CAM) [1]. The Internet provides

unprecedented opportunities to create powerful expert systems for biotechnology. In such systems, opportunities for obtaining information (OI) could be created rather cheaply as well as the best way to exchange large volumes of data to compile diverse, even contradictory, evidence and opinions from medical experts, different fields of medicine and remote geographic regions. Clinical decision support systems (CDSS) have become increasingly used in biomedical practice in the last decades. The first such system, which has been widely used since 1970, has become the medical expert system MYCIN. After it, a number of systems were created that provided access to medical information, interpretation of diagnoses, and so on. During development of these technical information systems (IS), one of the problem is choosing methods for efficient system construction and data usage from the databases for biotechnology. For such problem solution two groups of methods are used: automatic OI and receiving OI in manual mode (manual OI).

The method of automatic OI [1]. It is also called "knowledge discovery" and "data mining" — "knowledge acquisition" — is relatively new. The most important step of it is to extract abstract rules from a large number of cases. The most widely used automated OI methods are cluster analyzes, neural networks, discriminant and linear programming, evolutionary algorithms, and etc. But even with such perfect mathematic apparatus sometimes it is impossible to solve these problems due to extremely complex algorithms for mentioned methodologies. For example, when searching for data from large databases, some data may be correct or incorrect, and this will influence output data.

Manual OI [1]. As a result, most of the modern biotechnical knowledge bases refer to manual OI, although knowledge bases designed for this method are usually not large, referring to very specific and relatively narrow areas of medicine. Manually operated OIs are usually constructed in close collaboration with experts and engineers in biotechnology. For example in medicine, the development of manual OI takes a lot of time, and it is important that the medical diagnosis is a complex cognitive process that medical experts sometimes cannot formalize. Manual OI is not always available everywhere, therefore, not all users outside of biotechnical or health centers can use these systems.

The Internet can solve these problems better than traditional platforms, since 1 — the Internet is widely available; 2 — Web browsers provide a common multimedia interface; 3 — for the development of expert systems there is software that can be obtained from the Internet; 4 — there are protocols for support of the interaction between such expert systems; 5 — experts can communicate online through the Internet, eliminate duplication of information, and so on.

Application of cluster methods for data analysis in biotechnology. Application of cluster methods for databases and IS in modern researches is quite common [1–4, 12–16]. Methods of cluster analysis (MCA) are the most advanced mathematical methods, widely used in modern biology and medicine, and can be successfully used in biotechnology, for example, in expert systems, for automatic OI (see above). Descriptions of different examples of cluster analysis methods' use in biotechnology one can find in electronic publications [1, 2, 5, 6–16]. As was demonstrated above, cluster analysis methods could be applied successfully in all cases where it is necessary to distinguish between two or few different objects. Thus, Jézéquel et al. [6] used these methods for identification of three subtypes of triple-negative patients: luminal androgen receptor (22%), basal-like with low immune response and high M2-like macrophages (45%), and basal-enriched with high immune response and low M2-like macrophages (33%). They noted that macrophages and other immune effectors offer a variety of therapeutic targets in breast cancer, and particularly in triple-negative basal-like tumours. Furthermore, they demonstrated that CK5 antibody was better suited than CK5/6 antibody to subtype triple-negative patients.

In publication of Ko et. al. [8] methods of cluster analysis were used for studying how expression profiling of ion channel genes predicts clinical outcome in breast cancer. The authors identified a molecular gene signature IC30, which represents a promising diagnostic and prognostic biomarker in breast cancer. Their results indicate that information regarding the expression of ion channels in tumor pathology could provide new targets for therapy in human cancers.

In work of Kawai et al. [9] it was demonstrated using MCA that midostaurin preferentially attenuates proliferation of triple-negative breast cancer (TNBC) cell lines through inhibition of Aurora kinase family. There was shown that midostaurin suppresses the proliferation of TNBC cells among the breast cancer cell lines presumably through the inhibition of the Aurora kinase family.

The authors proved that the precise study of midostaurin on cell growth will contribute to the development of the drug for the treatment of TNBC.

Other authors [9] used these methods for analysis of the results of drug sensitivity screening for understanding drugs' effect in case of breast cancer. In total, 25 correlated and four anticorrelated drug sensitivities were revealed of which only one drug, Sirolimus, showed significantly lower IC50 values in the luminal/ERBB2 breast cancer subtype. The authors found the expected interactions and discovered new relationships between drugs which might have implications for cancer treatment as well.

A lot of contemporary investigations of genome and gene engineering works are carried out using cluster analysis, like clustering analysis of proteins from microbial genomes [13] or studying of EcoGene tools applied to the RefSeq prokaryotic genomes [14]. Application of cluster analysis method (Ward method) for early diagnosis of tumor processes is given in [7], where the practical application of the Ward method to detect, record and analyze differences in the complex of biochemical parameters of blood in transfected mice (teratocarcinoma T-36) and in healthy mice in control was demonstrated. This analysis was carried out using clustering method that allowed to divide all the examined mice by purely analytical indicators into 2 classes, one of which was healthy, and the second — mice with transplanted tumors. In turn, the last experimental class of mice was subdivided into 2 subclasses according to the terms after tumor transplantation: up to 30 days (the beginning of the death of animals) and with long survival times. Thus, it has been shown that standard laboratory blood parameters (hematologic and biochemical) make it possible to distinguish experimental groups of overweight mice already in the early stages of tumor development.

Cluster analysis (CA), used in all observed articles, is a multi-dimensional statistical procedure that collects data which contain information on object selection, then arranges them in relatively homogeneous groups. The clustering task refers to statistical processing.

CA methods should be used to solve such problems:

1. Understanding the data by identifying cluster structure. Splitting a sample into groups of similar objects allows ones to simplify further data processing and decision making, applying to each cluster method of analysis.

2. Data compression. If the initial sample is too large, one can reduce it by leaving one of the most typical representative of each cluster.

3. Novelty detection. Using cluster methods it is possible to find outstanding objects that cannot be linked to any of the clusters.

In all these cases, hierarchical clustering can be applied, when large clusters are split into smaller ones, those into smaller ones, etc. Such tasks are called taxonomy tasks, the result of which is a tree-like hierarchical structure. At the same time, each object is characterized by the enumeration of all clusters to which it belongs, usually from the largest to the smallest.

Input data types. Input data used for cluster analysis in biotechnology have to include the following:

1. Description of object's characteristics. Each object is described by a set of its characteristics, which are called signs. Signs can be numerical or non- numerical.

2. Matrix of distances between objects. Each object is described by distances to all other objects of metric space.

3. Matrix of similarity between objects. It is necessary to take into account the degree of similarity of an object with other sample objects in metric space. Similarity here complements the distance (difference) between objects to 1.

Some systems of clustering methods classification. There are several systems of clustering methods classification used in different countries. In fact, they include the same sets of clustering procedures, only based on different principles of their grouping in classificational hierarchy. Below we suggest their classification list drawn by scientists of Eastern European countries (Slavonic-speaking countries), further — by their colleagues from English-speaking countries (ESC).

1) *Classification of cluster methods in Eastern European countries*. Scientists there suppose that there is no one commonly used classification of cluster methods, but it is possible to distinguish some groups of approaches. Some methods are possible to associate with few groups at once, so the suggested version has to be seen only as an approximation to the real, more perfect future classification of clustering methods.

1. Probability approach. It is supposed that each studied object belongs to one of k classes. Some authors suppose that this group does not even belong to clustering and call it "discrimination": the choice of inclusion

of objects to one of known groups (training samples).

- k-means.
- k-medians.
- EM-algorithm.
- Algorithms of FOREL family.
- Discriminant analysis.

2. Approaches based on artificial intelligence. This is a group of wide notions because it includes very many methods and they are really different.

- Method of fuzzy clusterization (C-means).
- Kohonen neural network.
- Genetic algorithm.

3. Logical approach. Construction of dendrogramma is done based on the tree of solutions.

4. Theoretical and graph approach.

- Graph algorithms of clustering.

5. Hierarchical approach. Existence of included groups (clusters of different orders) is supposed. Further these algorithms are subdivided onto agglomerative (united) and divisive (splitted). According to the number of characteristics it is possible to subdivide monothetic and polythetic classification methods.

- Hieracrhical divisive clustering or taxonomy. Clustering tasks are studied in numerical taxonomy.

6. Other methods were not included into previous groups:

- Statistical algorithms of clustering.
- Assemble of clusterings.
- Algorithms of KRAB family.
- Algorithm on the basic method of sieving.
- DBSCAN and others.

Approaches 4 and 5 sometimes are united under the names "structural" or "geometric approach" because they have more formalized notion of proximity. Despite significant differences between the listed methods, all of them have got the same basic idea "hypothesis of compactness". This means that in space all near-located objects must belong to one cluster, and all different objects must belong to different clusters.

2) *Classification types of clustering methods in English-speaking countries.* A large number of clustering methods have been developed; the most widely used of them in ESC are classified as following.

1. By models of linkage: for example, using hierarchical clustering methods, models are elaborated based on the distance of the links (connected objects) (see above "Input data types").

2. Centroid models: for example, the algorithm "k-means" represents each cluster with a single vector of averages.

3. Distribution models: clusters are modeled using statistical methods of distribution, such as multivariant normal distribution using the "expectation-maximization" algorithm.

4. Density models: for example, DBSCAN and OPTICS define clusters, as connected segments of data space according to their density.

1) *The method of hierarchical clustering (or: a method based on linkages; linkage-based clustering, hierarchical clustering)* [1, 4] (Fig. 1). Using the CA method, one can make a cluster in some hierarchy. This clustering method based on a linkage is also known as hierarchical clustering. It is based on the basic idea of objects, one of which is more related to neighboring objects than more far objects. These algorithms connect "objects" with the formation of "clusters" based on their distance. The cluster can be described by such characteristics as the

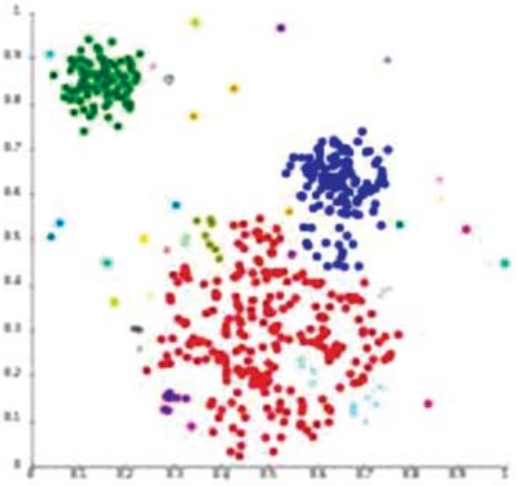

Fig. 1. **An example of a method of linkage clustering: single linkage on Gaussian data [2]**

maximal distance required for connecting parts of the cluster. At different distances various clusters are formed, they can be represented by a dendrogram (hence the name "hierarchical clustering"). These algorithms do not provide a simple (single) section of the data set but provide a hierarchy of clusters that merge with each other at certain distances. In the dendrogram the ***y*** axis represents the distance at which the cluster merges, while objects are placed along the ***x*** axis so that the clusters do not mix.

Linkage-based clustering is a family of methods that are characterized in how distances are calculated [1, 4]. For calculations, in addition to the usual choice of remote functions, it is also necessary to determine the criterion for linking of the distance. The most commonly used methods are: single linkage clustering for minimal distance to the object, complete linkage clustering for the maximal distance to the object, and *UPGMA* ("the method of non-calculating paired groups with arithmetic mean", or "medium links clustering"). In addition, the hierarchical clustering can be agglomerative (starting from individual elements and then uniting them into clusters), or separating (starting from the complete set of data and then dividing them into parts). This group also includes the Ward method from statistics, which is a criterion in a hierarchical cluster analysis [1, 4]. This method offers a general procedure for agglomerative hierarchical clustering, when the criterion for choosing of cluster pair fusion is based on the optimal value of object function.

This method of data clustering is not very convenient in case there are too many data too far from the cluster. Taking into account these "fluctuations" into a hierarchical cluster is difficult, and for a large set of data the calculation process takes a lot of time.

2) *The method of centroid-based clustering* (Fig. 2)

Several centroid algorithms for clustering sequences based on word counting are now being investigated [4]. An open source tool is implemented for clustering without alignment. The method allows clustering of sequences with high bandwidth, despite the limitations. In clusters based on centroids, they are represented by a central vector that does not necessarily be a member of the data set. When the number of clusters is fixed to k, the k-means clusters give an official definition as an optimization task: to find cluster centers and to assign objects to the closest cluster center, so that the square distances from the cluster are minimized. A particularly well-known approximation method is "k-means algorithm" (or: Lloyd's algorithm). Most k-means-type algorithms require a predetermined number of clusters, which is considered one of the major disadvantage of these algorithms. In addition, algorithms prefer clusters of approximately the same size, since they will always assign an object to the nearest centroid.

3) *The method of distribution-based clustering* (Fig. 3)

This model of clustering is most closely related to statistics [1, 4], based on distribution models. Clusters can be identified as objects that are the most likely to be similar. It is convenient that this approach is very similar to the way of creating artificial data sets: by selecting random objects from the distribution.

One of commonly used methods is known as the model of Gaussian mixes (using the

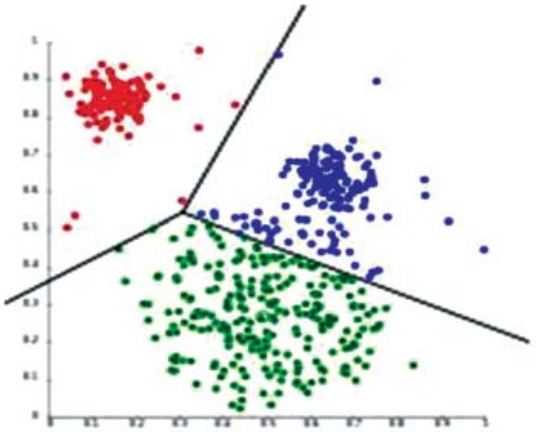

Fig. 2. **An example of a method of k-means clustering [2]**

maximization expectation algorithm). Here, the data set is usually modeled using a fixed (to avoid overflow) number of randomly initialized Gaussian distributions, whose parameters are iteratively optimized for better matching of the data set. This will coincide with the local optimum, so several starts can be followed by different results. To get the rough clustering, the objects later often are assigned to a Gaussian distribution; for soft clustering it is not necessary.

4) *The method of density-based clustering* (Fig. 4)

In a density-based method, clusters are defined as regions of greater density than the rest of the data set. Objects in these rarefied areas (needed to separate clusters) are usually considered as noise or border points. Clusters with similar density are evaluated; therefore some problems might appear with adjacent clusters separating.

The most popular method of clustering is DBSCAN, which has a well-defined cluster model called "density". The method is similar to clustering based on links (see type 1); it is based on points of links at certain distances. However, it unites only the points that satisfy the density criterion, in the original version, which is defined as the minimal number of other objects within this radius. A cluster consists of objects that are united by their density (which can form a cluster of any form, in contrast to many other methods) plus all objects that are within these objects. Another interesting property of DBSCAN is that its complexity is rather low — it requires a linear number of requests in the range of the database — and that it will really demonstrate the same results (this is determined for nucleus and noise points, but not for the boundary points). The average shift is a clustering, when each object moves to the densest nearest area,

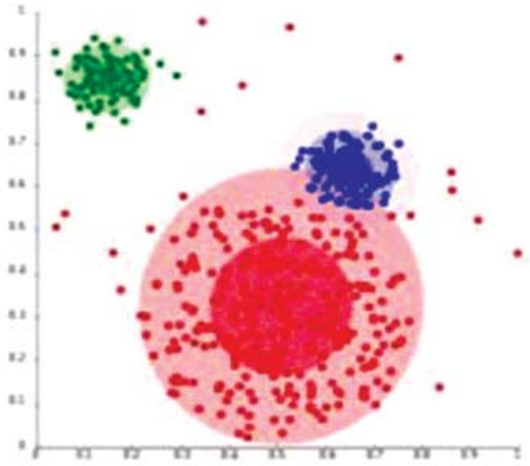

Fig. 3. **An example of a method of expectation-maximization [2]**
On Gaussian-distributed data, EM works well, since it uses Gaussians for modeling clusters

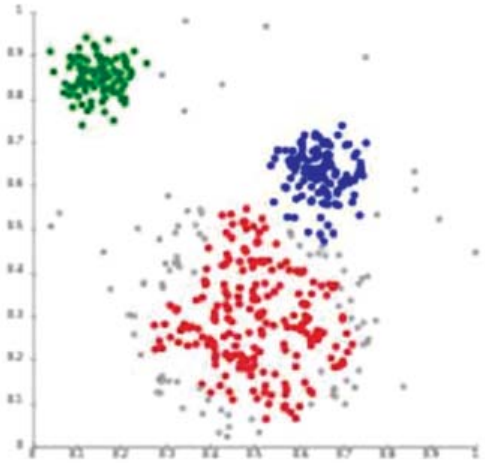

Fig. 4. **An example of a method of density-based clustering [2]**
Density-based clustering with DBSCAN

based on the estimation of nucleus density. Finally, the objects converge with the local maxima of density.

What clustering methods are the most suitable for the solution of medical and biological problems? Some attempts were done to determine which of clustering methods are the most suitable for solving problems in medicine and biology, for example in the publication of Iakovidis et al. [4]. Below we shall analyze such attempt following the author; further we shall observe some works that support the analysis that was done.

The feasibility of using cluster methods could be demonstrated under consideration for data sets with computer diagnosis of meningoencephalitis, during which there is a need to distinguish between similar signs of brain cells both healthy and infected by a virus. As we have already shown, the same problems of cells distinguishing are important for biotechnology as well. So, in modern medical practice we need to distinguish between the data concerning various types of pathogens (which are organisms with similar characteristics), manifestations of diseases with similar symptoms, etc. For example, at the first day of the disease even some doctors can make the wrong diagnosis — influenza, acute respiratory disease, but not the actual meningoencephalitis. Thus, if the patient is not provided immediate expert assistance, the lethal result is possible, and we have to observe the experience of succesful elaboration of a system for computer diagnostics of the disease, with an algorithm based on cluster methods. Thus, even without knowing the details of the domain database or some marks (e.g. names of diseases following doctors diagnosis), the cluster method can generate separate data. By such approaches, in some cases, knowing about these new generated classes, you can diagnose a new illness. Let's analyze what cluster methods used were the most successful in practice for medical data sets. For comparison it was selected and evaluated in terms of applications for data of meningoencephalitis four clustering techniques: agglomerative hierarchical cluster method with a single and complete clustering (single-, complete-linkage agglomerative hierarchical clustering), Ward method and rough clustering (see classification above). Below is a list of matched methods that were analyzed; it includes the following types of clustering methods.

1) *Agglomeration hierarchical clustering (AIC) method with single linkage.* It belongs to type 1 (ESC), namely hierarchical clustering, single linkage clustering.

2) *AIC with complete linkage* communications. It belongs to type 1 (ESC), namely hierarchical clustering, complete linkage clustering.

3) *Ward method.* It belongs to type 1 (ESC), namely hierarchical clustering, Ward method.

4) *Rough clustering*. It belongs to type 3 (ESC), namely distribution-based clustering, rough clusterization.

Experimenting with clusters, the authors compared the differences between theoretically generated clusters and classes diagnosed in practice. The appropriateness of considering certain measures of similarity was assessed on the basis of the following aspects:

1) the quality of the generated clusters;

2) whether the meanings in the clinic have the attributes used to generate high-quality clusters.

The results of the study of objects from the relevant databases, which contained 140 objects and their corresponding 32 attributes, demonstrated that using the method of Ward, the best clusters with attribute combinations that have meaning from the point of view of clinical practice were obtained.

For comparison, for each cluster method, a single similarity measure, a linear combination of the Mahalanobis distance between numerical attributes and the Hamming distance between nominal attributes were given. The utility of clustering methods was evaluated by the quality of the generated clusters, the correspondence between the attributes used to generate high-quality clusters and clinical experience. Below the logical course of provided analysis is given [4].

For comparison, for each cluster method, a single similarity measure, a linear combination of the Mahalanobis distance between numerical attributes and the Hamming distance between nominal attributes were given. The utility of clustering methods was evaluated by the quality of the generated clusters, the correspondence between the attributes used to generate high-quality clusters and clinical experience. Note that such methods should be used for medical datasets to elaborate the clinical databases.

In the publication [4] the possibility of elaboration of a database with information about meningoencephalitis was investigated. The four listed cluster methods were used to analyze datasets for diagnosis of meningoencephalitis, which contained 140 objects. The table shows the structure of data sets. Each object has 33 attributes, including

one attribute class and 32 other attributes. The attribute class and 12 of the 32 attributes are nominal, others are numeric. The DIAG attribute class shows the type of meningitis (bacterial or viral in origin). In this dataset, 2 of the 32 attributes are important for describing classes — counting polynuclear cells (Cell_Poly) and mononuclear cells (Cell_Mono). If polynuclear cells dominate, the patient is diagnosed with bacterial meningitis. If mononuclear cells are dominant, then they are diagnosed with viral meningitis.

Measures of similarity. Further the measures of similarity were studied. Let's assume that $U = \{x_1, x_2, \ldots, x_N\}$ is a set of objects, where N — is the number of objects, and each object has $p = p_c + p_d$ attributes, where p_c is the number of numerical attributes, and p_d is the same for nominal attributes. We define as an object $x_i = \{x_i^1, x_i^2, \ldots, x_i^p\}$, where x_i^j is the value of the j-th attribute of the object x_i.

Similarity for numeric attributes. In order to find similarity for numerical attributes, the modified expression for the Mahalanobis distance was used through a variant covariant matrix. If all attributes are independent and all attribute values are standardized, Mahalanobis distance for objects is equal to Euclidean distance:

$$d_{\mathrm{E}}(x_i,x_j) = \{(x_i^1 - x_j^1)^2 + (x_j^2 - x_j^2)^2 + \ldots + (x_i^p{}_c - x_j^p{}_c)\}^{1/2}.$$

Similarity for categorical attributes. In order to find similarity for categorical attributes, it was suggested to modify the Hamming distance to calculate the number of attributes for which two objects have different attribute values.

Similarity for mixed attributes. If objects have both numeric and categorical attributes, their similarity is calculated as the sum of the Mahalanobis distance weight $d_M(x_i,x_j)$ of the numerical attributes and the Hamming $d_H(x_i,x_j)$ of the nominal attributes.

Characteristics of some used cluster methods [1, 4].

Agglomeration hierarchical clustering. Hierarchical clustering (IC) has become widely used in cluster analysis, since it allows you to visualize the hierarchical structure of dendrogram clusters. For the most part, two types of algorithms are proposed for IC: agglomerate IC (AIC) and separate IC (DIC). In the case of AIC, an independent cluster is assigned to each object. Then in this method, we find such a pair of clusters and they merge into a single cluster. This process is repeated until all the original clusters merge into a single cluster. Separate hierarchical clusterization is a procedure opposite to AIC. It starts from one cluster and ends with a division into a certain number of clusters of an object. In cases of both methods, the merging or division ends when the step of merging/dividing of further clusters becomes too large. Agglomerative hierarchical clustering has some approaches that define strategies of cluster merging. Let's observe some of them.

Single linkage. The only way in which a cluster can be subdivided is to identify the differences between groups that exist for the closest pairs:

$$d_{SL}(G, H) = min_{x_i \in G,\, x_{i'} \in H}\, d(x_i,x_i'),$$

where G and H are clusters that merge in the next step. Clustering based on this distance is called agglomeration clustering with a single linkage, or the method of the closest neighbor.

Complete linkage. Another way to select a cluster is to identify the differences between groups for the case of the most remote pairs:

$$d_{CL}(G, H) = \max_{i \in G, i' \in H} d_{ii'},$$

where G and H are clusters that merge in the next step. Clustering based on this distance is called agglomerative clustering with complete linkages, or the method of the furthest neighbor.

Ward method. The Ward method is based on the sum of squares in the cluster. Let x_{li}^k and n_l denote the value of k-th attributes in i examples in cluster l and the number of examples in l. At the beginning the sum of the square of clusters l, S_l has to be determined. Now let's assume that the cluster l and the cluster m are integrated into the cluster lm and the sum of the cluster m, S_m and cluster lm is determined:

$$S_{lm} = S_l + S_m + \Delta S_{lm}.$$

Two clusters will integrate when ΔS_{lm} is the minimum for all clusters.

Clustering based on rough separation. In the general case, if the similarity of the objects is presented only as a relative similarity, then creating clusters that could be interpreted is difficult, since some important measures are difficult to define. A clustering method based on a rough separation is such that it allows clustering objects that do not differ. It allows you to give the entity of the objects according to the degree of indiscernibility and to make interpreted clusters even for the above mentioned objects. This method of clustering is based on the concept of indistinguishability of objects. Let us introduce some fundamental differences for clustering by a rough

An example of a meningoencephalitis dataset [4]

Object	Age	Sex	Feler	Focal	...	DIAG
1	10	M	10	+	...	BACTERIA
2	12	M	5	-	...	BACTERIA
. . .					. . .	
140	23	F	10	+	...	VIRUS

separation. Assume that $U \neq \varphi$ is a discourse universum, and X is a subset of U. The equivalence relation R classifies U as a set of subsets $U/R = \{X_1, X_2, ..., X_m\}$ in which the following conditions are satisfied:

1) $X_i \subseteq U$, $X_i \neq \phi$ for any i,
2) $X_i \cap X_j = \phi$ for any i, j,
3) $U_{i=1,2,...,n} X_i = U$.

Any subset X_i, called as a category, represents a class of equivalence R. Category in R contains an object $\in U$, denoted by $[x]_R$. For a family of equivalence relations $\boldsymbol{P} \subseteq \boldsymbol{R}$ the ratio of indistinguishability for $\boldsymbol{P}$ is denoted by IND($\boldsymbol{P}$) and it is determined as follows:

$$\text{IND}(\boldsymbol{P}) = \cap_{R \in P} \text{IND}(\boldsymbol{R}).$$

The clusterization method consists of two stages: 1) the establishment of initial equivalence relations; 2) iterative improvement of the initial relations of equivalence. At the first stage, initial equivalence ratios were established for each object. The initial equivalence relation classifies the object from the point of view of two sets: a set of objects similar to this object, and a set of objects that are not similar to this object. Let $U = \{x_1, x_2, ..., x_n\}$ be a complete set of n objects. Then the initial relation of the equivalence of R_i for objects x_i is defined as following:

$$R_i = \{\{P_i\}, \{U - P_i\}\};$$

$$P_i = \left\{ x_j \middle| s(x_i, x_j) \geq S_i \right\}, \ \forall x_j \in U$$

where P_i is a set of objects similar to x_i. For example, P_i is a set of objects whose similarity to x_i, s is higher than the threshold value S_i. The threshold value of S_i is determined automatically in the place where the value of s decreases the most strongly. A set of indistinguishable objects was obtained using the entire set of equivalence relations according to the cluster. In other words, the cluster corresponds to the category X_i out of $U/\text{IND}(\boldsymbol{R})$.

At the second stage, it is necessary to improve the initial equivalence ratios according to their general interrelations. Objects with a high degree of indistinguishability can be interpreted as similar objects. Therefore, they must be put into the same cluster. Thus, we modify the equivalence relation according to its ability to distinguish between objects with large γ in the following way:

$$R'_i = \{\{P'_i\}, \{U - P'_i\}\};$$

$$P'_i = \left\{ x_j \middle| \gamma(x_i, x_j) \geq T_h \right\}, \forall x_j \in U$$

This prevents the generation of small clusters that were formed in accordance with too deep detail during the classification T_h which is the threshold value that determines the indistinguishability of objects. Thus, we associate T_h with the inaccuracy of knowledge about the objects and perform the iterative perfection of the equivalence relations for the values of T_h, which are constantly decreasing. Accordingly, a set of rough-graded sequences is obtained as U/IND(R′).

Thus, at the beginning of this article we have shown the great significance that contemporary clustering methods have for expert and other computer system in biotechnology. The types of input data for tasks solution by such systems were described using methods of clasterization. Further the classification of clustering methods was given in comparison with Eastern European countries and English-speaking countries. In [4] it was discussed what kind of clustering method is the most suitable for the solution of medical and biological problems. For this purpose four methods were examined: three methods of hierarchical clustering (1 — single linkage clustering; 2 — complete linkage clustering; 3 — Ward method) and one method of distribution-based clustering (rough clustering). In the observed publication the possibility of elaboration of a database with information about meningoencephalitis was investigated.

So, like it was explained at the beginning of this article, demonstrated above as well as in publications [1–16], and numerous other publications [41–82], the clustering methods for data analysis could be used to solve some tasks in biotechnology too. High level of mathematical abstraction of these methods and existion of some similar features of biomedical objects permit us to apply these for biotechnology as

well. Besides, these methods are possible to use for determining whether there are two (or more) formations of the same object, or these are two different objects in terms of mathematics. Such powerful mathematic methods were developed for such problems solution for different areas of science and practice. Their application in biotechnology, like other biological and medical sciences, became so popular in the last decades, that we feel obliged to make a profound description of these methods' capabilities, to show a wide spectrum of these methods and to demonstrate practical application of some selected ones as the most successful for biomedical object analysis. Some other examples of cluster analysis methods used in biology and medicine were observed in this article [6–16, 48–82], because similar types of problems arise during the analysis of processes of grooving of cultural masses, cell differentiation, studying of genes [8] (Fig. 5), or tumor pathology [9] (Fig. 6), etc. In numerous publications [17–82], it was demonstrated that such methods are valuable in general for biomathematics and bioinformatics as well. Thus it is reasonable to use cluster analysis methods for computer diagnostics in cases when it is necessary to distinguish between cells with weak differences (in norm and pathology). Other examples of cluster analysis methods use for biomedical objects from [1, 2, 5–16] are also given above. The new idea [1, 11] is the application of cluster analysis methods for the solution of the task of elements distinguishing during the elaboration of electronic systems with databases that may be used for biotechnology purposes (see the described examples of technical systems CDSS, expert system MYCIN, technical information systems with automatic OI, manual OI; as well as DBSCAN, OPTICS, and etc.) [1, 14, 15].

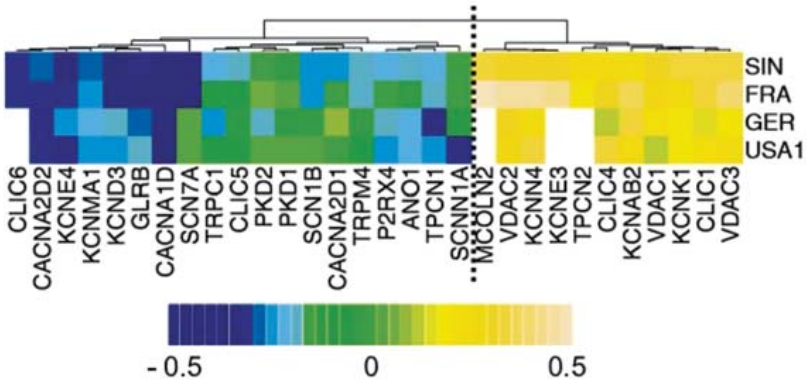

Fig. 5. **New targets for pathologies therapy revealed using cluster analysis for studying the expression of ion channels in tumor pathology [8]**

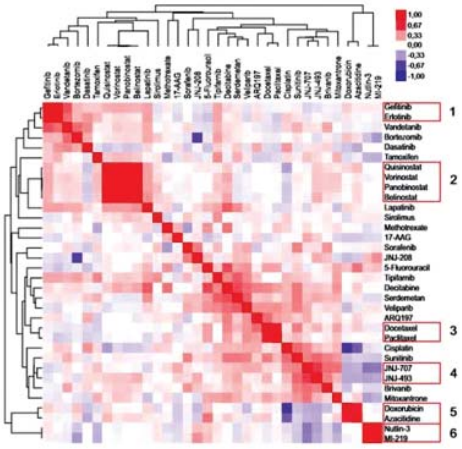

Fig. 6. **Examination of preparations for new schemes of therapy Pearson correlation plot of absolute drug IC_{50} values [10]:**
The red color indicates a positive correlation between the IC_{50} values of two drugs, and blue is a negative correlation. The color intensity illustrates the correlation coefficient as shown in the legend at top right. Drugs are clustered on the basis of similarity; distances in the tree indicate the degree of difference between drugs

REFERENCES

1. *Klyuchko O. M.* Information and computer technologies in biology and medicine. *Kyiv: NAU-druk.* 2008, 252 p. (In Ukrainian).
2. *Merrell R., Diaz D.* Comparison of data mining methods on different applications: clustering and classification methods. *Inf. Sci. Lett.* 2015, 4 (2), 61–66. http://dx.doi.org/10.12785/isl/040202.
3. *Jecheva V., Nikolova E.* Some clustering-based methodology applications to anomaly intrusion detection systems. *Int. J. Secur. Appl.* 2016, 10 (1), 215–228. http://dx.doi.org/10.14257/ijsia.2016.10.1.20.
4. *Iakovidis D. K., Maroulis D. E., Karkanis S. A.* Texture multichannel measurements for cancer precursors' identification using support vector machines. *Measurement.* 2004, V. 36, P. 297–313. https://doi.org/10.1016/j.measurement. 2004.09.010
5. *Nguyen H. Q., Carrieri-Kohlman V., Rankin S. H., Slaughter R, Stulbarg M. S.* Internet-based patient education and support interventions: a review of evaluation studies and directions for future research. *Comp. Biol. Med.* 2004, 34 (2), 95–112. doi: 10.1016/S0010-4825(03)00046-5.
6. *Jézéquel P., Loussouarn L., Guérin-Charbonnel C., Campion L., Vanier A., Gouraud W., Lasla H., Guette C., Valo I., Verrièle V., Campone M.* Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Res.* 2015, 17 (1), 43. https://doi.org/10.1186/s13058-015-0550-y.
7. *Bozhenko V. K.* Multivariable analysis of laboratory blood parameters for obtaining diagnostic information in experimental and clinical oncology. The dissertation author's abstract on scientific degree editions. *Dc. Med. Study. Moscow.* 2004. (In Russian).
8. *Ko J. H., Ko E. A., Gu W., Lim I., Bang H., Zhou T.* Expression profiling of ion channel genes predicts clinical outcome in breast cancer. *Mol. Cancer.* 2013, 12 (1), 106. doi: 10.1186/1476-4598-12-106.
9. *Kawai M., Nakashima A., Kamada S., Kikkawa U.* Midostaurin preferentially attenuates proliferation of triple-negative breast cancer cell lines through inhibition of Aurora kinase family. *J. Bbiomed. Sci.* 2015, 22 (1), 48. doi: 10.1186/s12929-015-0150-2.
10. *Uhr K., Wendy J. C., Prager-van der Smissen, Anouk A. J. Heine, Bahar Ozturk, Marcel Smid, Hinrich W. H. Göhlmann, Agnes Jager, John A. Foekens, John W. M. Martens.* Understanding drugs in breast cancer through drug sensitivity screening. *SpringerPlus.* 2015, 4 (1), 611. doi: 10.1186/s40064-015-1406-8.
11. *Onopchuk Yu. M., Biloshitsky P. V., Klyuchko O. M.* Development of mathematical models based on the results of researches of Ukrainian scientists at Elbrus. *Visnyk NAU.* 2008, N 3, P. 146–155. (In Ukrainian).
12. *Ankur Poudel, Dhruba Bahadur Thapa, Manoj Sapkota.* Cluster Analysis of Wheat (*Triticum aestivum* L.) Genotypes Based Upon Response to Terminal Heat Stress. *Int. J. Appl. Sci. Biotechnol.* 2017, 5 (2), 188–193. doi: http://dx.doi.org/10.3126/ijasbt.v5i2.17614.
13. *Zaslavsky L., Ciufo S., Fedorov B., Tatusova T.* Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC Bioinform.* 2016, 17 (8), 276. Published online 2016 Aug 31. doi: 10.1186/s12859-016-1112-8.
14. *Zhou J., Richardson A. J., Rudd K. E.* EcoGene-RefSeq: EcoGene tools applied to the RefSeq prokaryotic genomes. *Bioinformatics.* 2013, 29 (15), 1917–1918. Published: 04 June 2013. doi: 10.1093/bioinformatics/btt302.
15. *Zhang J., Wu G., Hu X., Li S., Hao S.* A Parallel Clustering Algorithm with MPI — MKmeans. *J. Comput.* 2013, 8 (1), 10–17. doi: 10.1109/PAAP. 2011.17.
16. *Tatusova T., Zaslavsky L., Fedorov B., Haddad D., Vatsan A., Ako-adjei D., Blinkova O., Ghazal H.* Protein Clusters. *The NCBI Handbook [Internet]. 2nd edition.* Available at https://www.ncbi.nlm.nih.gov/books/NBK242632.
17. *Anderson J. G.* Evaluation in health informatics: computer simulation. *Computers in Biology and Medicine.* 2002, 32 (3), 151–164. https://doi.org/10.1016/S0010-4825(02)00012-4.
18. *Aruna P., Puviarasan N., Palaniappan B.* An investigation of neuro-fuzzy systems in psychosomatic disorders. *Exp. Syst. Appl.* 2005, 28 (4), 673–679. https://doi.org/10.1016/j.eswa.2004.12.024.
19. *Baert P., Meesen G., De Schynkel S., Poffijn A., Oostveldt P. V.* Simultaneous in situ profiling of DNA lesion endpoints based on image *cytometry* and a single cell database approach. *Micron.* 2005, 36 (4), 321–330. https://doi.org/ 10.1016/j.micron.2005.01.005.
20. *Bange M. P., Deutscher S. A., Larsen D., Linsley D., Whiteside S.* A handheld decision support system to facilitate improved insect pest management in Australian cotton systems. *Comp. Electron. Agricult.* 2004, 43 (2), 131–147. https://doi.org/10.1016/j.compag.2003.12.003.
21. *Beaulieu A.* From brainbank to database: the informational turn in the study of the brain. *Stud. Hist. Phil. Biol. Biomed. Sci.* 2004, V. 35, P. 367–390. https://doi.org/10.1016/j.shpsc.2004.03.011.
22. *Bedathur S. J., Haritsa J. R., Sen U. S.* The building of BODHI, a bio-diversity database system. *Inform. Syst.* 2003, 28 (4), 347–367. https://doi.org/ 10.1016/S0306-4379(02)00073-X.

23. *Berks G., Ghassemi A., von Keyserlingk D. G.* Spatial registration of digital brain atlases based on fuzzy set theory. *Comp. Med. Imag. Graph.* 2001, 25 (1), 1–10. https://doi.org/10.1016/S0895-6111(00)00038-0.
24. *Brake I.* Unifying revisionary taxonomy: insect exemplar groups. *Abstr. XV SEL Congr. Berlin (Germany).* 2007.
25. *Braxton S. M., Onstad D. W., Dockter D. E., Giordano R., Larsson R., Humber R. A.* Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL. *J. Invertebr. Pathol.* 2003, 83 (3), 185–195. doi: 10.1016/S0022-2011(03)00089-2.
26. *Breaux A., Cochrane S., Evens J., Martindaled M., Pavlike B., Suera L., Benner D.* Wetland ecological and compliance assessments in the San Francisco Bay Region, California, USA. *J. Environm. Manag.* 2005, 74 (3), 217–237.
27. *Budura A., hilippeCudré-Mauroux P., Aberer K.* From bioinformatic web portals to semantically integrated Data Grid networks. *Future Generation Computer Systems.* 2007, 23 (3), 281–522. doi: 10.1016/j.jenvman.2004.08.017.
28. *Burns G., Stephan K. E., Ludäscher B., Gupta A., Kötter R.* Towards a federated neuroscientific knowledge management system using brain atlases. *Neurocomputing.* 2001, V. 38–40, P. 1633–1641. https://doi.org/10.1016/S0925-2312(01)00520-3.
29. *Butenko S., Wilhelm W. E.* Clique-detection models in computational biochemistry and genomics. *Eur. J. Oper. Res.* 2006, 173 (1), 1–17. https://doi.org/ 10.1016/j.ejor.2005.05.026.
30. *Carro S. A., Scharcanski J.* Framework for medical visual information exchange on the WEB. *Comp. Biol. Med.* 2006, 36 (4), 327–338. doi: 10.1016/ j.compbiomed.2004.10.004.
31. *Chaplot S., Patnaik L. M., Jagannathan N. R.* Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. *Biomed. Signal Process. Control.* 2006, 1 (1), 86–92. https://doi.org/10.1016/j.bspc.2006.05.002.
32. *Chakravarty M. M., Bertrand G., Hodge C. P., Sadikot A. F., Collins D. L.* The creation of a brain atlas for image guided neurosurgery using serial histological data. *NeuroImage.* 2006, 30 (2), 359–376. doi: 10.1016/ j.neuroimage.2005.09.041.
33. *Chau M., Huang Z., Qin J., Zhou Y., Chen H.* Building a scientific knowledge web portal: The NanoPort experience. *Decision Support Systems.* 2006. https://doi.org/10.1016/j.dss.2006.01.004.
34. *Chen M., Hofestädt R.* A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modeling, and systematic analysis. *J. Biomed. Inform.* 2006, 39 (2), 147–159. https://doi.org/10.1016/j.jbi.2005.05.005.
35. *Chli M., De Wilde P.* Internet search: Subdivision-based interactive query expansion and the soft semantic web Applied Soft Computing. 2006. https://doi.org/10.1016/j.asoc.2005.11.003.
36. *Despont-Gros C., Mueller H., Lovis C.* Evaluating user interactions with clinical information systems: A model based on human–computer interaction models. *J. Biomed. Inform.* 2005, 38 (3), 244–255. https://doi.org/10.1016/ j.jbi.2004.12.004.
37. *Despont-Gros C., Mueller H., Lovis C.* Evaluating user interactions with clinical information systems: a model based on human-computer interaction models. *J. Biomed. Inform.* 2005, 38 (3), 244–255. doi: 10.1016/j.jbi.2004.12.004.
38. *Marios D., Dikaiakos M. D.* Intermediary infrastructures for the World Wide Web. *Comp. Networks.* 2004, V. 45, P. 421–447. https://doi.org/ 10.1016/j.comnet.2004. 02.008.
39. *Dikshit A., Wu D., Wu C., Zhao W.* An online interactive simulation system for medical imaging education. *Comp. Med. Imag. Graph.* 2005, 29 (6), 395–404. https://doi.org/10.1016/j.compmedimag.2005.02.001.
40. *Dimitrov S. D., Mekenyan O. G, Sinks G. D., Schultz T. W.* Global modeling of narcotic chemicals: ciliate and fish toxicity. *J. Mol. Struc.: Theochem.* 2003, 622 (1–2), 63–70. https://doi.org/10.1016/S0166-1280(02)00618-8.
41. *Dong Y., Zhuang Y., Chen K., Tai X.* A hierarchical clustering algorithm based on fuzzy graph connectedness. *Fuzzy Sets. Syst.* 2006, V. 157, P. 1760–1774. https://doi.org/10.1016/j.fss.2006.01.001.
42. *Duan Y., Edwards J. S., Xu M. X.* Web-based expert systems: benefits and challenges. *Inf. Manag.* 2005, 42 (6), 799–811. https://doi.org/10.1016/j.im.2004. 08.005.
43. *Essen van D. C.* Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr. Opin. Neurobiol.* 2002, 12 (5), 574–579. https://doi.org/10.1016/S0959-4388(02)00361-6.
44. *Fellbaum C., Hahn U., Smith B.* Towards new information resources for public health — From Word Net to Medical Word Net. *J. Biomed. Inform.* 2006, 39 (3), 321-332. doi: 10.1016/j.jbi.2005.09.004.
45. *Ferraris M., Frixione P., Squarcia S.* Network oriented radiological and medical archive. *Comp. Physics Commun.* 2001, V. 140, P. 226–232. https://doi.org/10.1016/S0010-4655(01)00273-9.
46. *Flower D. R., Attwood T. K.* Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors. *Semin. Cell. Dev. Biol.* 2004, 15 (6), 693–701. doi: 10.1016/j.semcdb.2004.09.008.
47. *Fink E., Kokku P. K., Nikiforou S., Hall L. O., Goldgof D. B, Krischer J. P.* Selection of patients for clinical trials: an interactive web-based system. *Art. Intell. Med.* 2004, 31 (3), 241–254. doi: 10.1016/j.artmed.2004.01.017.

48. *Fitzpatrick M. J., Ben-Shahar Y., Smid H. M., Vet L. E., Robinson G. E., Sokolowski M. B.* Candidate genes for behavioural ecology. *Trend Ecol. Evol.* 2005, 20 (2), 96–104. doi: 10.1016/j.tree.2004.11.017.
49. *Fox J., Alabassi A., Patkar V., Rose T., Black E.* An ontological approach to modelling tasks and goals. *Comp. Biol. Med.* 2006, V. 36, P. 837–856. https://doi.org/10.1016/j.compbiomed.2005.04.011.
50. *Fu Zetian, Xu Feng, Zhou Yun, Shuan X. Z.* Pig-vet: a web-based expert system for pig disease diagnosis. 2006. https://doi.org/10.1016/j.eswa.2005.01.011.
51. *Gaulton A., Attwood T. K.* Bioinformatics approaches for the classification of G-protein-coupled receptors. *Curr. Opin. Pharmacol.* 2003, 3 (2), 114–120. doi: 10.1016/S1471-4892(03)00005-5.
52. *Gevrey M., Worner S., Kasabov N., Pitt J., Giraudel J. L.* Estimating risk of events using SOM models: A case study on invasive species establishment. *Ecol. Modell.* 2006, 197 (3–4), 361–372. https://doi.org/10.1016/j.ecolmodel. 2006.03.032.
53. *Glenisson P., Glänzel W., Janssens F., Moor B. D.* Combining full text and bibliometric information in mapping scientific disciplines. *Inf. Proc. Manag.* 2005, 41 (6), 1548–1572. https://doi.org/10.1016/j.ipm.2005.03.021.
54. *Gomez-Perez A., Fernandez-Lopez M., Corcho O.* Ontological engineering. *London: Springer-Verlag.* 2004. https://doi.org/10.1007/b97353.
55. *Graham C. H., Ferrier S., Huettman F., Moritz C., Peterson A. T.* New developments in museum-based informatics and applications in biodiversity analysis. *Trend. Ecol. Evol.* 2004, 19 (9), 497–503. https://doi.org/10.1016/ j.tree.2004.07.006.
56. *Gruber T. R.* A translation approach to portable ontologies. *Knowledge Acquisition*. 1993, 5 (2), 199–220. doi: 10.1006/knac.1993.1008.
57. *Hauser C., Holstein J., Steiner A.* Butterfly taxonomy for the Internet: opportunities and challenges for the GART/GloBIS database project. *Abst. XIV SEL Congress. Roma (Italy).* 2005, P. 21.
58. *Hirano S., Sun X., Tsumoto S.* Comparison of clustering methods for clinical databases. *Inform. Sci.* 2004, 159 (3–4), P. 155–165. https://doi.org/ 10.1016/j.ins.2003.03.011.
59. *Hong Yu., Hatzivassiloglou V., Rzhetsky A., Wilbur W. J.* Automatically identifying gene/protein terms in MEDLINE abstracts. *J. Biomed. Inform.* 2002, 35 (5–6), 322–330. https://doi.org/10.1016/S1532-0464(03)00032-7.
60. *Horn W.* AI in medicine on its way from knowledge-intensive to data-intensive systems. *Artificial Intelligence in Medicine. Elsevier.* 2001, 23 (1), 5–12. https://doi.org/10.1016/S0933-3657(01)00072-0.
61. *Hsi-Chieh Lee, Szu-Wei Huang, Li E. Y.* Mining protein–protein interaction information on the internet. *Exp. Syst. Appl. Elsevier.* 2006, 30 (1), 142–148. https://doi.org/10.1016/j.eswa.2005.09.083.
62. *Jabs R., Pivneva T., Huttmann K., Wyczynski A., Nolte C., Kettenmann H., Steinhäuser C.* Synaptic transmission onto hyppocampal glial cells with hGFAP promoter activity. *J. Cell Sci.* 2005, V. 118, P. 3791–3803. doi: 10.1242/jcs.02515.
63. *Johnson S. B., Friedman R.* Bridging the gap between biological and clinical informatics in a graduate training program. *J. Biomed. Inform.* 2007, 40 (1), 59–66. Epub. 2006 Mar 15. doi: 10.1016/j.jbi.2006.02.011.
64. *Kaiser M., Hilgetag C. C.* Modelling the development of cortical systems networks. *Neurocomputing.* 2004, V. 58–60, P. 297–302. https://doi.org/ 10.1016/j.neucom.2004.01.059.
65. *Kane M. D., Brewer J. L.* An information technology emphasis in biomedical informatics education. *J. Biomed. Inform.* 2007, 40 (1), 67–72. https://doi.org/10.1016/j.jbi.2006.02.006.
66. *Kannathal N., Acharya U. R., Lim C. M., Sadasivan P. K.* Characterization of EEG — A comparative study. *Comp. Meth. Progr. Biomed.* 2005, 80 (1), 17–23. https://doi.org/10.1016/j.cmpb.2005.06.005.
67. *Kitching I. J.* Taxonomy in the 21 Century: the CATE model for web revisions. *Abst. of XV SEL Congress. Berlin (Germany).* 2007.
68. *Koh W., McCormick B. H.* Brain microstructure database system: an exoskeleton to 3D reconstruction and modelling. *Neurocomputing.* 2002, V. 44–46, P. 1099–1105. https://doi.org/10.1016/S0925-2312(02)00426-5.
69. *Koh W., McCormick B. H.* Registration of a 3D mouse brain atlas with brain microstructure data. *Neurocomputing*. 2003, V. 52–54, P. 307–312. https://doi.org/10.1016/S0925-2312(02)00793-2.
70. *Kovalev V. A., Petrou M., Suckling J.* Detection of structural differences between the brains of schizophrenic patients and controls. *Psychiatry Research: Neuroimaging.* 2003, 124 (3), 177–189. https://doi.org/10.1016/S0925-4927(03) 00070-2.
71. *Kulish V., Sourin A., Sourina O.* Human electro encephalograms seen as fractal time series: Mathematical analysis and visualization. *Comp. Biol. Med.* 2006, 36 (3), 291–302. doi: 10.1016/j.compbiomed.2004.12.003.
72. *Li Q., Wu Y.* Identifying important concepts from medical documents. *J. Biom. Inform.* 2006, 39 (6), 668–679. doi: 10.1016/j.jbi.2006.02.001.
73. *Lubitz von D., Wickramasinghe N.* Networkcentric healthcare and bioinformatics: Unified operations within three domains of knowledge. *Exp. Syst. Appl.* 2006, 30 (1), 11–23. https://doi.org/10.1016/j.eswa.2005.09.069.

74. *Ma Y., Hof P. R., Grant S. C., Blackband S. J., Bennett R., Slatest L., McGuigan M. D., Benveniste H.* A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Neuroscience*. 2005, 135 (4), 1203–1215. doi: 10.1016/j.neuroscience.2005.07.014.
75. *Mahaman B. D., Harizanis P., Filis I., Antonopoulou E., Yialouris C. P., Sideridis A. B.* A diagnostic expert system for honeybee pests. *Comp. Electr. Agricult*. 2002, 36 (1), 17–31. https://doi.org/10.1016/S0168-1699(02)00069-8.
76. *Martin-Sanchez F., Iakovidis I., Nørager S., Maojo V., de Groen P., Van der Lei J., Jones T., Abraham-Fuchs K., Apweiler R., Babic A., Baud R., Breton V., Cinquin P., Doupi P., Dugas M., Eils R., Engelbrecht R., Ghazal P., Jehenson P., Kulikowski C., Lampe K., De Moor G., Orphanoudakis S., Rossing N., Sarachan B., Sousa A., Spekowius G., Thireos G., Zahlmann G., Zvárová J., Hermosilla I., Vicente F. J.* Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J. Biomed. Inform.* 2004, 37 (1), 30–42. doi:10.1016/j.jbi.2003.09.003.
77. *Masseroli M., Visconti A., Bano S. G. Pinciroli F.* HealthCo-op: a web-based system to support distributed healthcare co-operative work. *Comp. Biol. Med.* 2006, 36 (2), 109–127. doi:10.1016/j.compbiomed.2004.09.005.
78. *Moon S., Byun Y., Han K.* FSDB: A frameshift signal database. *Comp. Biol. Chem.* 2007, 31 (4), 298–302. doi: 10.1016/j.compbiolchem.2007.05.004.
79. *Nowinski W. L., Belov D.* The Cerefy Neuroradiology Atlas: a Talairach–Tournoux atlas-based tool for analysis of neuroimages available over the internet. *NeuroImage.* 2003, 20 (1), 50–57. https://doi.org/10.1016/S1053-8119(03)00252-0.
80. *Orgun B., Vu J.* HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems. *Comp. Biol. Med.* 2006, 36 (7–8), 817–836. https://doi.org/10.1016/j.compbiomed.2005.04.010.
81. *Pérez-Rey D., Maojo V., García-Remesal M., Alonso-Calvo R., Billhardt H., Martin-Sánchez F., Sousa A.* Ontofusion: Ontology-based integration of genomic and clinical databases. *Comp. Biol. Med.* 2006, 36 (7–8), 712–730. doi: 10.1016/j.compbiomed.2005.02.004.
82. *Rana B. K., Insel P. A.* G-protein-coupled receptor websites. *Trend. Pharmacol. Sci.* 2002, 23 (11), 535–536. doi: http://dx.doi.org/10.1016/S0165-6147(02)02113-2.

КЛАСТЕРНИЙ АНАЛІЗ У БІОТЕХНОЛОГІЇ

О. М. Ключко

Інститут експериментальної патології, онкології та радіобіології ім. Р.Є.Кавецького НАН України, Київ

E-mail: kelenaXX@ukr.net

Метою роботи був опис методів кластерного аналізу та доказ можливості їх застосування в біотехнології. Оскільки є певний досвід застосування цих методів у біології та медицині, проаналізовано відповідні публікації. Наведено коротку характеристику основних принципів кластерного аналізу, їх використання в біології та медицині та окремі приклади — в біотехнології. Розглянуто результати вирішення біотехнологічних проблем за допомогою кластерних методів у комплексі з іншими математичними підходами. У висновках наведено результати виконаного аналізу, а також рекомендації щодо використання методів кластерного аналізу в біотехнології.

Ключові слова: кластерний аналіз, біотехнологія.

КЛАСТЕРНЫЙ АНАЛИЗ В БИОТЕХНОЛОГИИ

Е. М. Ключко

Институт экспериментальной патологии, онкологии и радиобиологии им. Р. Е .Кавецкого НАН Украины, Киев

E-mail: kelenaXX@ukr.net

Целью работы был анализ методов кластерного анализа и возможность их применения в биотехнологии. Поскольку существует определенный опыт применения этих методов в биологии и медицине, проанализированы соответствующие публикации. Приведена краткая характеристика основных принципов кластерного анализа, использование их в биологии и медицине и отдельные примеры — в биотехнологии. Рассмотрены результаты исследований биотехнологических проблем с помощью кластерных методов в комплексе с другими математическими подходами. В выводах приведены результаты выполненного анализа, а также рекомендации по применению методов кластерного анализа в биотехнологии.

Ключевые слова: кластерный анализ, биотехнология.