Applying biclustering technique and gene ontology analysis for gene expression data processing

Sergii Babichev[1,2,*,†], Maksym Korobchynskyi[3,†], Myhailo Rudenko[3,†] and Hanna Batenko [1,†]

[1] *Kherson State University, University street, 27, Kherson, 73000, Ukraine*
[2] *Jan Evangelista Purkyne University in Usti nad Labem, Pasteurova, 15, Usti nad Labem, 400 96, Czech Republic* [3] *Military Academy named after Eugene Bereznyak, Yria Il'enka street, 81, Kyiv, 04050, Ukraine*

Abstract

This study details the biclustering methods for gene expression data, focusing on the refinement of quality criteria essential for evaluating the generated bicluster structures. An internal biclustering quality criterion is introduced, leveraging mutual information evaluation across both rows and columns within a bicluster. Additionally, the research proposes a novel hybrid biclustering model, which amalgamates the ensemble biclustering algorithm with Bayesian optimization techniques to optimize the algorithm's parameters effectively. This model is grounded on a target objective function derived from the newly proposed quality criterion. Simulations carried out on gene expression data from patients afflicted with various cancer types demonstrate the efficacy of the model. Specifically, the application of the mutual information-based criterion within the objective function results in the formation of a bicluster structure comprising 18 distinct biclusters. Furthermore, the study expands upon a method that employs gene ontology analysis, facilitating the identification of subsets of significant gene expression data from bicluster analysis results. A comprehensive procedure for identifying significant gene subsets through a combination of bicluster and gene ontology analyses is executed. The evaluation of sample classification results, characterized by these significant gene subsets, underscores the method's effectiveness. The classification quality criteria exhibit relatively high values, even with a reduced number of genes, indicating the method's efficiency.

Keywords

gene expression data, bicluster analysis, gene ontology analysis, biclustering quality criteria, convolution neural network (CNN)

1. Introduction

The significance of bicluster analysis in the processing of gene expression data is determined by its possibility to allocate the subsets of mutually coherent gene expression values, which can improve the effectiveness of disease diagnosis systems [1,2]. Unlike traditional clustering

IntelITSIS'2024: 5th International Workshop on Intelligent Information Technologies and Systems of Information Security, March 28, 2024, Khmelnytskyi, Ukraine

* Corresponding author.
† These authors contributed equally.
sergii.babichev@ujep.cz (S.Babichev); maks_kor@ukr.net (M. Korobchynskyi); ruminik33@ukr.net (M. Rudenko); gbatenko@ksu.ks.ua (H. Batenko)
0000-0001-6797-1467 (S.Babichev); 0000-0001-8049-4730 (M. Korobchynskyi); 0000-0002-9180-1510 (M. Rudenko); 0000-0001-7007-4708 (H. Batenko)

techniques, which primarily concentrate on categorizing objects (either rows or columns) according to their likeness, thus neglecting the potential significance of interactions among various data dimensions, biclustering introduces a more comprehensive strategy. When applying the bicluster analysis, the concurrent grouping of both rows and columns is performed, thereby allocating subsets of data that exhibit mutual correlations. This capability is particularly crucial when dealing with intricate datasets, as it facilitates a deeper understanding of the underlying patterns and relationships. Biclustering methods stand out by offering a dual-axis analysis framework, which is fundamental to dissecting complex biological data, such as gene expression profiles. This analysis not only identifies clusters of genes with similar expression patterns across a subset of conditions but also pinpoints conditions under which these genes exhibit similar behavior. The dual grouping mechanism inherent in bicluster analysis is indispensable for exploring and interpreting the multifaceted nature of gene expression data, revealing insights into gene functions, regulatory mechanisms, and cellular processes that might remain obscured under traditional clustering approaches.

In the context of information technology and bioinformatics, ontology is a formalized version of knowledge representation that utilizes a controlled vocabulary and a set of relationships between terms to describe the domain being considered [3,4]. Such ontology can be used for modeling the subject area and serve for information exchange, data integration, and the development of various computer applications, including artificial intelligence. In bioinformatics, ontologies are used to structure and standardize information about biological processes, protein functions, cellular components, and more. The Gene Ontology (GO) is an example of such a system that allows for the annotation of genes and protein products in a unified form, ensuring consistency and compatibility of biological databases. Biclustering and data analysis based on gene ontology are linked through their common goal - understanding the biological mechanisms and functional characteristics of genes that are revealed in experimental gene expression data. While biclustering allows identifying groups of genes that show similar expression patterns under different conditions or in different types of samples (in the presence of different types of diseases), which is important for understanding which genes are co-regulated in certain physiological states or respond to certain external stimuli, data analysis in biclusters based on gene ontology allows determining the possible role of the highlighted genes in the cell or organism being studied. In other words, gene ontology represents a functional annotation of genes. Integrating biclustering results with analysis based on GO offers the possibility to gain a deeper understanding of the biological context of gene expression patterns and highlight groups of genes that are co-expressed in the presence of a certain type of disease. Thus, biclustering and analysis based on GO complement each other, providing a mechanism for the identification and functional understanding of biological modules in large sets of gene expression data. This fact indicates the relevance of the research topic.

2. Related works

The application of bicluster analysis for processing complex data has been the focus of a significant number of scientific works nowadays. For instance, [1] presents a review of metaheuristic approaches to solving biclustering problems, which effectively address complex optimization tasks within a limited computational time and adapt to various problem

formulations. Special attention is given to optimization methods and key search elements: representation, objective function, and variation operators, with a discussion on single- and multi-objective approaches and highlighting new research directions. In [2], the hidden block structure in a heterogeneous panel data model is explored, based on the assumption that regression coefficients have group structures among individuals and structural changes over time, where change points can affect group structures, and structural changes can vary between subgroups. To recover the hidden block structure, the authors propose a robust bicluster approach that uses M-estimation and concave penalty fusion, as well as developing an algorithm based on local quadratic approximation for optimizing the objective function, which is more compact and efficient compared to the ADMM algorithm. Furthermore, an oracle property for penalized M-estimators is established, and it is proven that the proposed estimator recovers the hidden block structure with high probability, which is also confirmed by positive results in practice through simulation studies on several datasets.

In [5], to improve the quality of biclustering and module extraction, a combination of methods based on Adaptive Resonance Theory (ART) is utilized - Biclustering ARTMAP (BARTMAP) and Topological ART (TopoART), which together form TopoBARTMAP. This method inherits the ability to detect topological associations while reducing data volume. TopoBARTMAP was tested on 35 real cancer datasets and compared with other (bi)clustering methods, showing statistically significant improvements over other evaluated methods in experiments with ordered and shuffled data. It also demonstrated better results in identifying constant, additive, multiplicative, and multiplicative-additive biclusters in experiments with 12 synthetic datasets. Graphical representation was refined to display associations of gene biclusters and evaluated on the NCBI GSE89116 dataset, which contains expression levels of 39,326 probes selected over 38 observations. In [6], a new biclustering algorithm for binary data called the Binary Biclustering Algorithm Based on Adjacency Difference Matrix (AMBB) was proposed, improving the balance between execution time and efficiency. The AMBB algorithm constructs an adjacency matrix based on adjacency difference values, and the resultant submatrix, updated using the adjacency difference matrix, is referred to as a bicluster. This allows for grouping genes that exhibit similar reactions under different conditions, which is important for further gene analysis. Experiments on synthetic and real datasets visually demonstrate the high practicality of the AMBB algorithm.

Despite the significant advancements in the field of bicluster analysis for processing complex data, there remain unresolved challenges, including the lack of effective methods for optimizing the hyperparameters of the relevant algorithm. This issue is particularly pertinent in the context of new approaches, such as the combination of methods based on Adaptive Resonance Theory for biclustering, which require precise tuning of hyperparameters for efficient operation. Additionally, there is the problem of balancing between execution time and algorithm efficiency, especially in situations involving binary data, where the development of new optimization strategies is needed to ensure fast and accurate data processing.

The **goal of the paper** is the development and application of the technique of gene expression data processing based on the joint use of bicluster analysis and gene ontology analysis methods.

3. Material and methods

3.1. Biclustering quality criterion based on an assessment of mutual information

As mentioned earlier, biclustering is the process of simultaneously clustering rows and columns of a matrix. In the context of gene expression data analysis, experimental data are represented as a matrix, where rows correspond to genes and columns to experimental conditions or vice versa, and the values in the matrix reflect the level of gene expression under a certain condition, i.e., its expression. In this case, a bicluster identifies a subset of genes that exhibit similar expression profiles across a subset of conditions. One way to assess the quality of a bicluster is through the application of mutual information (MI) analysis between rows and columns. MI can indicate how much the information in the rows and columns depends on each other, and thus, a high MI value may indicate a high coherence of the bicluster. The most common methods for estimating mutual information include the following [7,8]:

- Mutual Information (MI):

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

where: X, Y are vectors between which the MI is assessed; p(x, y) is the joint probability distribution of X and Y; p(x) and p(y) are the marginal probability distributions.

- Normalized mutual information is defined as the ratio of mutual information to the geometric mean of the entropies of the two vectors:

$$NMI(X,Y) = \frac{MI(X,Y)}{\sqrt{H(X),H(Y)}} \quad (2)$$

where H(X) and H(Y) are the entropies of vectors X and Y, respectively.

- Relative entropy, or Kullback-Leibler divergence, is a measure of the distance between two probability distributions:

$$D_{KL}(P||Q) = \sum_i P(i)log\frac{P(i)}{Q(i)} \quad (3)$$

where P(i) is the probability of distribution P, and Q(i) is the probability of distribution Q.

It should be noted that this distance is not symmetric, i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, hence, to enhance objectivity, it is advisable to calculate the bidirectional divergence with subsequent averaging of the two divergences:

$$D_{KL}(P,Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \quad (4)$$

Mutual information is a measure of shared information between two vectors of random variables, but it is not in itself a distance metric. Transforming the value of mutual information into a distance can be achieved in various ways. Within the scope of our research, a metric based on Shannon entropy is applied:

$$d(X,Y) = H(X) + H(Y) - 2MI(X,Y) \quad (5)$$

where H(X) and H(Y) are the Shannon entropy values of vectors X and Y, respectively, and MI(X,Y) is the mutual information between vectors X and Y. In this case, if considering two identical data distributions, then $H(X) = H(Y) = MI(X,Y)$ and $d(X,Y) = 0$. As the difference between data distributions increases, the value of mutual information decreases, leading to an increase in the distance between these vectors.

Calculating the value of the internal criterion for assessing the coherence of a bicluster involves estimating the average distance both between the rows and between the columns of the bicluster. The step-by-step procedure for calculating this criterion includes the following steps:

1. Calculation of the average distance among all pairs of rows in the bicluster:

$$QC_{row} = \frac{2}{nrow \times (nrow - 1)} \sum_{i=1}^{nrow-1} \sum_{j=i+1}^{nrow} d(X_i, X_j) \quad (6)$$

2. Calculation of the average distance among all pairs of columns in the bicluster:

$$QC_{col} = \frac{2}{ncol \times (ncol - 1)} \sum_{i=1}^{ncol-1} \sum_{j=i+1}^{ncol} d(Y_i, Y_j) \quad (7)$$

3. Calculation of the average value of the criteria (6) and (7):

$$QC = \frac{QC_{row} + QC_{col}}{2} \quad (8)$$

The minimum value of criterion (8) corresponds to the maximum level of bicluster coherence. It should be noted that when applying any clustering algorithm to gene expression data, characterized by a large volume of data, a fairly large number of biclusters with low coherence may emerge, which do not allow for a definitive identification of the class of samples under investigation. Moreover, the architecture of biclustering is largely determined by the parameters of the algorithm used to form the cluster structure. Therefore, the problem of optimizing algorithm parameters also arises, for which a Bayesian optimization algorithm is used within the current research, entailing the following stages:

Stage I. Definition of the objective function.

1.1. Selection of the biclustering algorithm that takes the values of the objective function parameters as input. Application of the algorithm to gene expression data. Formation of the bicluster structure.

1.2. Selection of a bicluster and assessment of its coherence using formulas (6) – (8).

Stage II. Definition of the parameter change range.

2.1. Determination of the range of variation for each parameter's values.

Stage III. Selection of the model and launch of the optimization algorithm.

3.1. Selection of the Bayesian optimization algorithm model. A model based on Gaussian processes was used in the research.

3.2. Application of the Bayesian optimization algorithm using the chosen model. Formation of the best combination of hyperparameters according to the formulated objective function.

Stage IV. Verification of the result and formation of a compromise decision regarding the optimal combination of hyperparameters.

4.1. Application of the aforementioned procedure to the first five biclusters (the number of biclusters may vary during modeling) followed by an analysis of the obtained results to form a compromise decision regarding the optimal combination of algorithm parameters.

Stage V. Application of the biclustering algorithm to gene expression data. Formation of the bicluster structure. Assessment of the coherence of the identified biclusters and formation of a subset of biclusters with a high coherence value for further research.

3.2. Biclustering quality criterion based on an assessment of mutual information

The procedure for identifying significant genes based on gene ontology analysis involves the use of the functions from the Bioconductor module [15,16] The practical implementation of this procedure assumes the following steps:

1. Loading necessary packages in R. During the simulation process, for the analysis of gene ontology and the selection of informative genes, the following packages were used: GO.db [17], org.Hs.eg.db [18], biomaRt [19], and topGO [20].

2. Data preparation. Creation of a list of gene identifier vectors (ENTREZ ID) contained in the identified biclusters.

3. Mapping genes to GO terms using functions from the org.Hs.eg.db package. Retrieval of GO terms for all genes contained in the bicluster.

4. Statistical analysis of gene expression values to estimate the probability (p-value) that the differences in gene expression values corresponding to different classes to which the samples belong could have occurred by chance. At this stage, the ANOVA (Analysis of Variance) statistical method was used, which allows comparing the mean values of three or more groups. In the context of gene expression analysis, ANOVA is used to determine whether there is a statistically significant difference in gene expression levels between different sample classes. The obtained p-values in this case indicate the probability that the observed differences could have occurred by chance. To adjust p-values (calculation of p-adjust), the Benjamini-Hochberg (BH) method was used to control type I errors during multiple comparisons.

5. Creating a topGO data object, which contains all gene identifiers and their scores, GO annotations, the hierarchical structure of GO, and all other information necessary for analyzing gene enrichment being studied.

6. Performing enrichment tests. Within the framework of dissertation research, two types of statistical tests were applied: the Fisher's exact test, based on counting the number of genes corresponding to each sample class, and the Kolmogorov-Smirnov test, which calculates enrichment based on assessments of gene expression values. Each of these tests provides an estimate of how differentially expressed a particular gene is, allowing genes to be categorized by their level of differential expression.

7. Formation of a gene ontology analysis result matrix with identifiers of genes that correspond to significant gene ontologies as a result of the analysis.

8. Formation of a vector of significant genes for the respective bicluster by finding matches between gene identifiers contained in the bicluster and gene identifiers identified as a result of the gene ontology analysis.

4. Experiment, results and discussion

4.1. Modeling to determine the optimal hyperparameters of the “ensemble” biclustering algorithm using the Bayesian optimization method

At this stage of modeling, gene expression data from patients studied for various types of cancer diseases were used as experimental data. The data are freely available on the website of The Cancer Genome Atlas project – TCGA [9] and contained nine sample classes, eight of which correspond to different types of cancer diseases, and the ninth group of gene expression data corresponds to subjects for whom cancer disease was not detected. Initially, the data contained 3269 samples and 19947 genes. After removing non-expressed and low-expressed genes for all samples using the method presented in [10], the number of genes was reduced to 19265. In the next step, mutually expressed gene expression profiles were identified from the obtained data by applying the inductive spectral clustering algorithm according to the method presented in [10], resulting in 3444 genes contained in the third cluster of the three-cluster structure (corresponding to the highest accuracy of sample classification). Thus, the experimental data had the form (3269×3444).

The modeling process was carried out in the R software environment [11] using the *biclust* package [12], which contains functions for applying various biclustering algorithms. Considering the studies presented in [13], within the current research, the biclustering process was performed using the *ensemble* algorithm [14], whose effectiveness, according to the results presented in [13], is significantly higher compared to using other biclustering algorithms. The outcome of the ensemble algorithm is determined by two parameters: the thresholding coefficient (thr) and the approximate ratio of the number of rows to columns in biclusters (simthr). The modeling process involved varying the values of these parameters within a predefined range with a certain step, calculating the values of the criterion (8) at each step of this procedure implementation. During the simulation procedure implementation, at each iteration, the first five biclusters were allocated, for each of which the value of the criterion was calculated. The evaluation of the biclustering was based on the average arithmetic value of all components of the corresponding criterion, which determines the coherence level of each of the identified biclusters. The analysis of the obtained results allowed us to conclude that the maximum value of the objective function (negative value of the criterion (8)) is achieved at the 10th iteration, with the following values of the "ensemble" biclustering algorithm parameters obtained: $thr = 0.549$; $simthr = 0.151$.

Table 1 shows the results of the “ensemble” biclustering algorithm with optimal hyperparameters values operation. As it can be seen, 18 biclusters various sizes were allocate in this case. The next stage is the application of gene ontology analysis to the data in the allocated biclusters.

4.2. Forming a subset of significant gene expression data using gene ontology analysis

The simulation process regarding the use of the gene ontology method to form a vector of significant genes, considering the type of samples, was carried out using gene expression data from the first bicluster. Figure 1 illustrates the result of applying the ANOVA statistical test to gene expression data (Volcano plot). The horizontal axis (Log2 Fold Change) on the diagram

shows the level of expression value of one group of genes compared to the expression of genes from another group. To the left of the center are genes that have lower expression in the first group compared to the second. Genes depicted to the right of the center have higher expression in the first group. It is evident that the further a gene is located from the center, the higher its level of differential expression. The vertical axis displays p-values (p-adjust) in a logarithmic scale (-log10(p-adjust)). Genes with lower p-values, indicating greater statistical significance of the difference in expression, are positioned higher on the graph. The analysis of the obtained results allows us to conclude that a relatively large number of genes contained in the bicluster can be identified as insignificant (located at the center bottom of the diagram), which confirms the need for further analysis with the aim of their removal.

Table 1

The result of the biclust analysis of gene expression data when applying ther “ensemble” algorithm

BC 1				BC 2				BC 3			
Gene	Sample			Gene	Sample			Gene	Sample		
456	gbm	lgg	norm	430	gbm	lgg	norm	399	luad	lusc	stad
	53	494	4		137	524	5		7	5	49
BC 4				BC 5				BC 8		BC 9	
Gene	Sample			Gene	Sample			Gene	Sample	Gene	Sample
339	luad	lusc		123	luad	lusc	stad	484	lusc	782	kirc	norm
	5	23			105	32	22		4		197	1
BC 6				BC 7				BC 11			
Gene	Sample			Gene	Sample			Gene	Sample		
189	luad	lusc	stad	329	luad	lusc	stad	463	luad	lusc	stad
	127	56	60		104	33	23		68	62	1
BC 10		BC 12			BC 14				BC 16		
Gene	Sample	Gene	Sample		Gene	Sample			Gene	Sample	
461	sarc	505	acc	sarc	348	luad	lusc	norm	612	kirc	norm
	13		5	2		14	1	4		299	1
BC 13		BC 15				BC 17			BC 18		
Gene	Sample	Gene	Sample			Gene	Sample		Gene	Sample	
450	lusc	315	luad	lusc	stad	346	acc	sarc	231	acc	sarc
	14		6	3	69		6	5		8	30

The next step involves performing enrichment tests with the calculation of p-values, which determine the level of significance of genes according to the respective test. As mentioned earlier, the Fisher's test and the Kolmogorov-Smirnov test were applied during the modeling process. The results of the modeling are shown in Figure 2.

According to both tests, 388 significant GO terms were identified out of 704. In the depicted diagram, the size of a dot is proportional to the number of annotated genes for the corresponding GO term, and its color represents the number of significantly differentially expressed genes. The thresholding parameter, which separates genes into significant and insignificant, was chosen at the level of the median of the gene significance vector. As can be seen, red dots contain many more genes than blue ones.

The analysis of the diagram presented in Figure 2 also allows us to conclude that the results of applying the Fisher's test and the Kolmogorov-Smirnov test differ from each other. Thus,

some GO terms identified as significant using the Fisher's test have less significance when applying the Kolmogorov-Smirnov test.

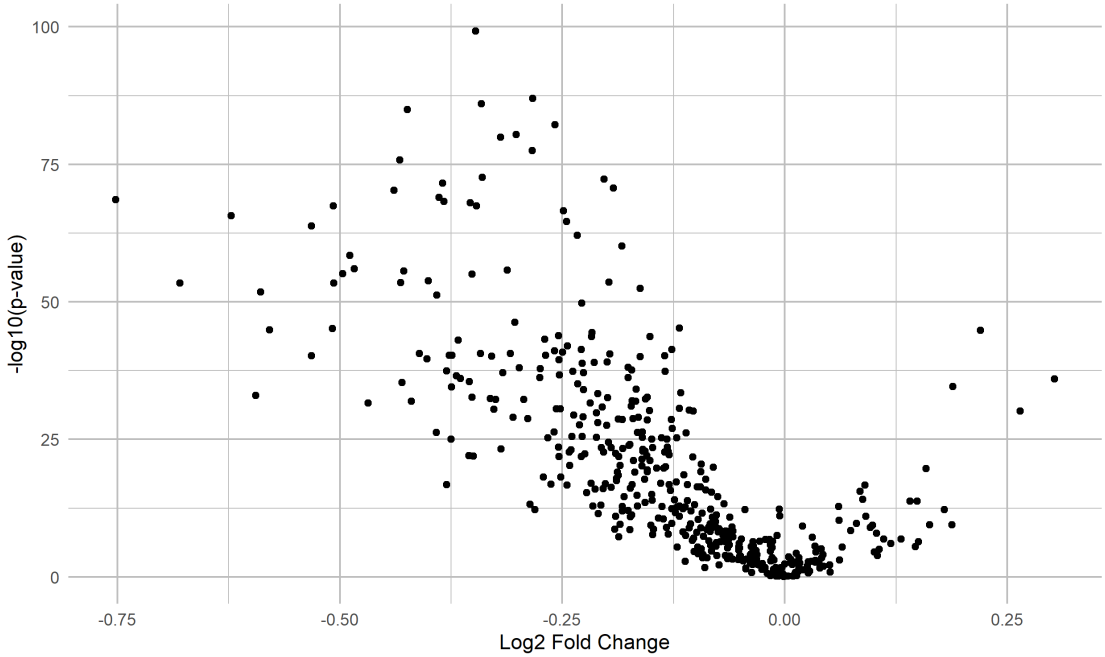

Figure 1: Visualization of the distribution of p-values of genes based on their significance level (volcano plot).

However, in some cases, it is possible to visually identify several GO terms for which the p-values are nearly the same when both tests are applied. The obtained results also indicate that despite the same number of significant genes when applying both tests, using only one test to form a subset of significant genes based on GO analysis is not objective. In this case, increasing the objectivity of the analysis can be achieved by applying both tests with the formation of intermediate decisions followed by their combination to select unique identifiers of significant genes.

In Figures 3 and 4, the results of applying GO analysis with the identification of ten significant GO terms using the Fisher's and Kolmogorov-Smirnov tests, respectively, are presented. Significant nodes are represented as rectangles. The color of the node represents relative significance, varying from dark red (most significant) to bright yellow (least significant). The analysis of the obtained graphs confirms the conclusion regarding the inconsistency of results when applying different tests during GO analysis aimed at forming a subset of significant genes. As evident from the figures, when identifying the ten most significant GO terms, the results differ both in the topology of the graph and in the significance level of the GO terms, which serve as the nodes of the graph. This fact corroborates the hypothesis about the advisability of applying both tests for forming a subset of significant genes.

As the simulation results have shown, the outcome of applying GO analysis is a table of convergence between GO terms and gene identifiers corresponding to the respective terms. Here, a single GO term can correspond to a large number of genes.

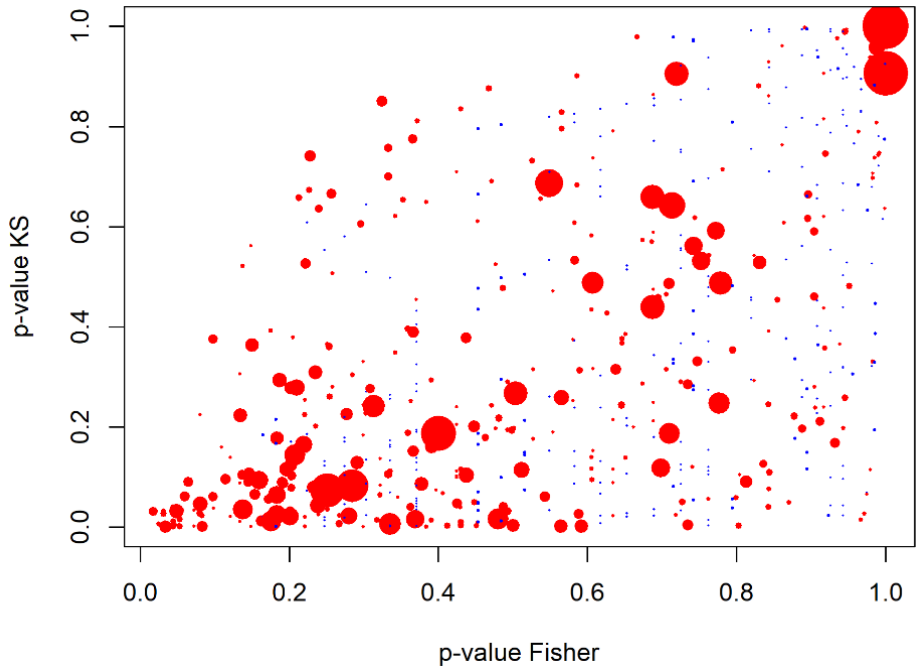

Figure 2: Scatter plot of the distribution of p-values calculated using the classical Fisher's test (x-axis) and the Kolmogorov-Smirnov method (y-axis).

For instance, in the case of applying the Fisher's test, the total number of genes corresponding to 388 significant GO terms was 26,092, whereas for the Kolmogorov-Smirnov test, it was 24,456.

In line with the set objective, the final step involved associating gene identifiers contained in the bicluster with gene identifiers identified through GO analysis. As a result, 270 genes were identified using the Fisher's test and 254 genes were identified using the Kolmogorov-Smirnov test. The total number of genes in the bicluster at this point was 465. By combining the results of applying both tests and identifying unique gene identifiers, the total number of significant genes amounted to 296.

However, it should be noted that the above type of GO analysis is effective when applied to data containing at least two sample classes, with each class having a sufficiently large number of samples. If these conditions are not met, the ANOVA test either will not work or may provide unreliable results.

For this reason, this type of GO analysis is suitably applied in the implementation of cluster analysis of gene expression profiles, where each cluster corresponds to a complete set of sample classes with a sufficiently large number of samples in each class.

When applying bicluster analysis, the condition for using the ANOVA test may not be met, as this can identify biclusters containing only one sample class, or the number of samples corresponding to one of the classes may be quite small, which reduces the reliability of the test results. In this case, it is appropriate to apply a statistical test based on the assessment of whether the number of genes associated with a certain GO term in the list of genes comprising the bicluster differs from the number expected by chance. In other words, the statistical test

compares the number of genes in the selected GO category contained in the bicluster with their total number in the genome of the studied object.

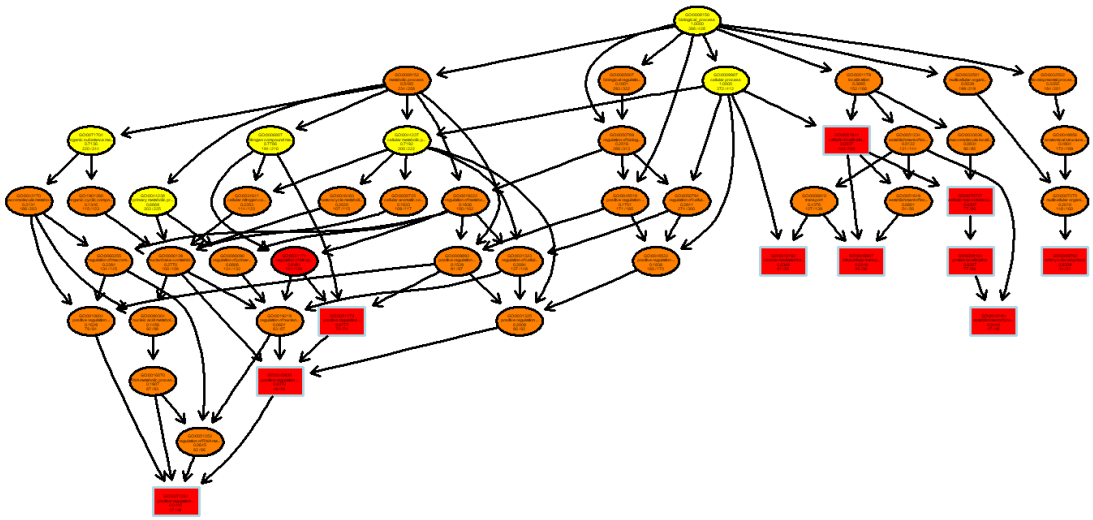

Figure 3: The result of applying GO analysis with the identification of ten significant GO terms using the Fisher's test.

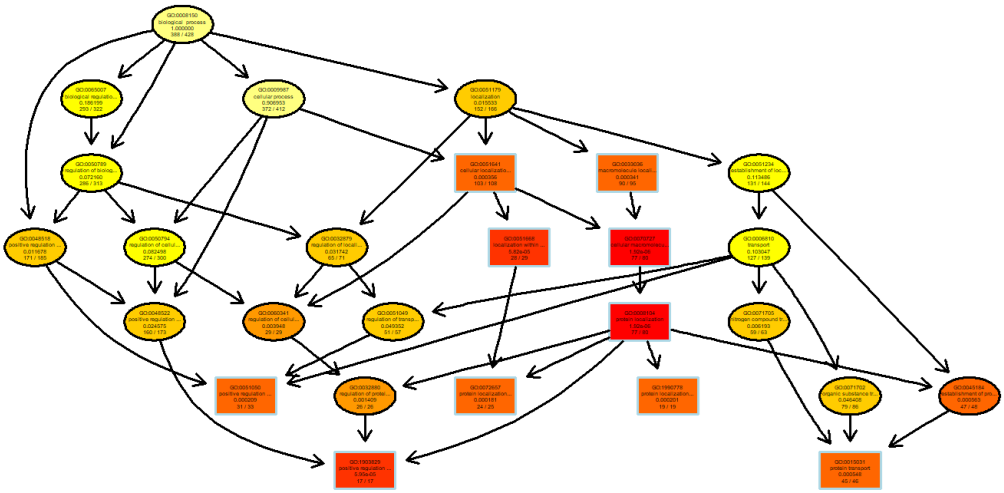

Figure 4: The result of applying GO analysis with the identification of ten significant GO terms using the Kolmogorov-Smirnov test.

Within the framework of dissertation research, the statistical test is implemented in the R programming environment by applying the function enrichGO() from the clusterProfiler package [21]. The use of the statistical test using the enrichGO() function involves two stages:

- Implementing the hypergeometric test by comparing the number of genes associated with a certain GO term to what is expected by chance. It should be noted that the GO term database must correspond to the type of biological object being studied. The GO term database for Homo sapiens “org.Hs.eg.db” was used in the modeling process.

- p-value correction. This step is necessitated by the large number of GO terms being analyzed, which requires adjusting p-values to control for multiple comparisons. The application of the Benjamini-Hochberg (BH) method helps to reduce the type I error.

The result of modeling regarding the application of GO analysis based on the enrichGO() function is a table with GO terms, which also contains p-values, adjusted p-values, and the number of genes in each term. Table 2 presents the result of the GO analysis of gene expression data from the first bicluster (the first 10 rows are shown). The threshold value that separates significant and insignificant GO terms was set at 0.05. At this threshold, 118 significant GO terms were identified. Figure 5 depicts the network of connections of the five most significant GO terms and their corresponding genes. As can be seen, as in the previous modeling, each GO term corresponds to a large number of genes, which confirms the need for filtering gene identifiers at a certain stage of data processing. Based on the modeling results regarding the application of gene ontology analysis to the bicluster structure formed in the previous step, which included 3444 genes, 1780 significant genes were identified. Thus, as a result of applying GO analysis, a new gene expression data matrix was formed: (3269×1780).

Table 2
Results of GO analysis using the statistical test based on the enrichGO() function with the application of gene expression data from the first bicluster

№	ID	GeneRatio	p-value	p.adjust	Count
1	GO:0007409	42/428	6.361398e-15	2.358170e-11	42
2	GO:0010975	41/428	4.477555e-14	8.299147e-11	41
3	GO:0050771	13/428	2.408776e-09	2.813517e-06	13
4	GO:0050770	19/428	3.035896e-09	2.813517e-06	19
5	GO:0050890	26/428	1.789554e-08	9.847434e-06	26
6	GO:0007411	22/428	1.859510e-08	9.847434e-06	22
7	GO:0097485	22/428	1.859510e-08	9.847434e-06	22
8	GO:0031345	19/428	8.314768e-08	3.852856e-05	19
9	GO:0048675	15/428	1.020838e-07	4.204719e-05	15
10	GO:0010977	16/428	1.213019e-07	4.496662e-05	16
…	…	…	…	…	…

To assess the effectiveness of the proposed technology, a one-dimensional two-layer GRU recurrent neural network was applied to the obtained data. The optimal number of neurons in the layers was determined using the Bayesian optimization algorithm. The number of neurons varied in the range from 20 to 100. The modeling results showed that increasing the number of neurons is not advisable, as it led to overfitting of the network. The discrepancy in classification results between the data used for training and model validation increased. According to the results of the Bayesian optimization algorithm, the number of neurons in the first and second layers was 98 and 75, respectively. Following the classical classifier application methodology, at the first stage, the data (samples) were divided into two subsets in a ratio of 0.7/0.3 (training subset and testing subset). At the second stage, the training subset was also divided into two subsets in a ratio of 0.8/0.2. The smaller subset was used for model validation during the training process. At each stage of applying the Bayesian optimization algorithm during network

training, 10-fold cross-validation was applied. Table 3 presents the classification results of samples that make up the testing subset of significant gene expression data.

Table 3

The results of classifying the samples comprising the test subset of expression data from significant genes selected using the criterion based on the MI evaluation

Class	Classification quality criteria				Total number	Correctly identified
	Precision	Recall	F1-score	Accuracy		
acc	0.893	0.893	0.893	0.943	28	25
gbm	0.915	0.915	0.915		59	54
kirc	0.982	0.964	0.973		169	163
luad	0.982	0.975	0.979		166	162
lgg	0.873	0.926	0.899		149	138
lusc	0.923	0.889	0.906		135	120
normal	0.908	0.952	0.929		62	59
sarc	0.962	0.949	0.955		79	75
stad	0.977	0.963	0.970		134	129

The analysis of the obtained results allows us to conclude that based on the group of classification quality criteria, the formation of bicluster structures followed by data filtering through the application of gene ontology analysis allows the formation of subsets of significant and mutually correlated gene expression data. The classification quality criteria values are consistently high in all cases, despite the limited number of genes that comprised the experimental data at the initial stage.

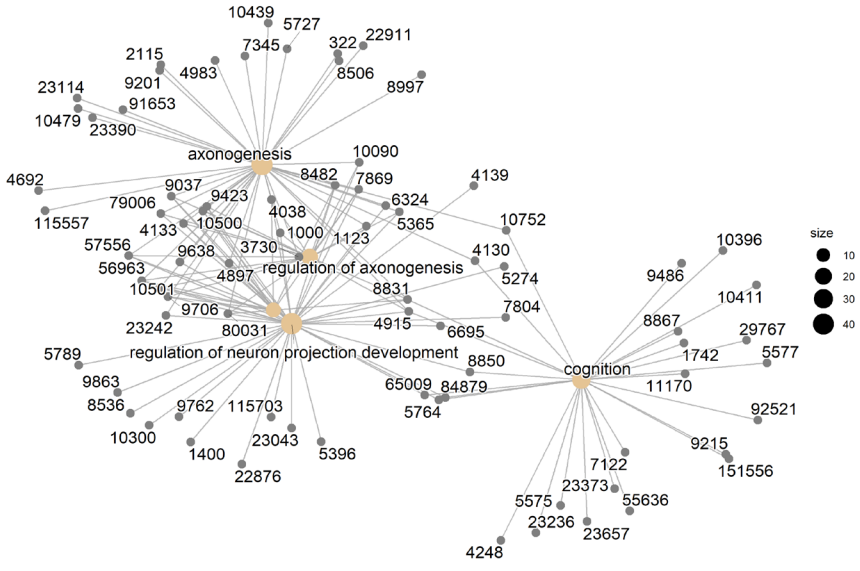

Figure 5: Network of connections of the five most significant GO terms with their corresponding genes.

5. Conclusions

This study presents the research results regarding the improvement of the methods of biclustering gene expression data by refining the quality criteria for biclustering, which allows us to evaluate the bicluster structure generated during the biclustering algorithm's execution. A novel internal quality criterion based on mutual information evaluation, both among the rows and columns of a bicluster, was proposed. Furthermore, a hybrid biclustering model for gene expression data processing has been proposed, integrating the ensemble biclustering algorithm and the Bayesian optimization method to fine-tune biclustering algorithm parameters. This model employs a target objective function based on the proposed quality criterion. Simulations using gene expression data from patients with various types of cancer showed that the objective function's application, using a criterion based on mutual information evaluation, formed the bicluster structure with 18 biclusters.

In this study also further developed a method based on gene ontology analysis in models, allowing for the formation of a subset of significant gene expression data using the results of the bicluster analysis.

We proposed and implemented the stepwise procedure for forming subsets of significant genes through the joint use of bicluster analysis and gene ontology analysis. The classification results, obtained using allocated significant gene expression data, underscored the methodological precision, with high scores across various metrics: precision, recall, F1-score, and accuracy, the values of which are varied within the range from 0.873 to 0.982, 0.889 to 0.975, 0.899 to 0.979, respectively, with an overall accuracy of 0.943.

The obtained results not only affirm the effectiveness of the joint use of biclustering and gene ontology analysis but also highlight the potential of applying deep neural network models to processing complex biological data.

The further prospects of the authors' research are the application of the proposed method within the framework of hybrid models of gene expression data processing based on the joint use of cluster-bicluster analysis, gene ontology analysis and deep learning methods.

References

[1] A. José-García, J. Jacques, V. Sobanski, C. Dhaenens. Metaheuristic Biclustering Algorithms: From State-of-the-art to Future Opportunities. ACM Computing Surveys, 2023, vol. 56(3), art. no. 69. doi: 10.1145/3617590

[2] W. Cui, Y. Li. Bicluster Analysis of Heterogeneous Panel Data via M-Estimation. Mathematics, 2023, vol. 11(10), art. no. 2333. doi: 10.3390/math11102333

[3] P.D. Tomas, D. Ebert, A. Muruganujan, et al. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Science, 2021, vol. 31, pp. 8-22. doi: 10.1002/pro.4218J. Cohen (Ed.), Special issue: Digital Libraries, volume 39, 1996.

[4] Ye. Hnatchuk, T. Hovorushchenko. Rules and method of supporting the decision-making regarding the possibility of extracorporeal fertilization. Computer Systems and Information Technologies, 2022, №3, pp. 6-10.

[5] R. Yelugam, L.E. Brito da Silva, D.C. Wunsch. Topological biclustering ARTMAP for identifying within bicluster relationships. Neural Networks, 2023, vol. 160, pp. 34-49. doi: 10.1016/j.neunet.2022.12.010

[6] H.-M. Chu, J.-X. Liu, K. Zhang, C.-H. Zheng, J. Wang, X.-Z Kong. A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. BMC Bioinformatics, 2022, vol. 23(1), art. no. 381. doi: 10.1186/s12859-022-04842-4.

[7] S. Babichev, J. Krejci, J. Bicanek, V. Lytvynenko. Gene expression sequences clustering based on the internal and external clustering quality criteria. Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, 2017, vol.1, art. no. 8098744, pp. 91-94. doi**:** 10.1109/STC-CSIT.2017.8098744.

[8] P. Schürger, V. Engel. Differential Shannon Entropies Characterizing Electron–Nuclear Dynamics and Correlation: Momentum-Space Versus Coordinate-Space Wave Packet Motion. Entropy, 2023, vol. 25(7), art. no. 970. doi**:** 10.3390/e25070970.

[9] The Cancer Genome Atlas Program. Available on: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

[10] S. Babichev, L. Yasinska-Damri, I. Liakh, J. Škvor. Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network. Applied Sciences (Switzerland), 2022, vol. 12(22), art. no. 11795. doi: 10.3390/app122211795.

[11] The R Project for Statistical Computing. Available on: https://www.r-project.org/

[12] S. Kaiser, R. Santamaria, T. Khamiakova, et al. Biclust: Bicluster Algorithms. Available on: https://cran.r-project.org/web/packages/biclust/index.html.

[13] S. Babichev, V. Osypenko, V. Lytvynenko, M. Voronenko, M. Korobchynskyi. Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. (2018), IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 - Proceedings, 2018, art. no. 8477439, pp. 298-304. doi: 10.1109/ELNANO.2018.8477439.

[14] P. Foszner, W. Labaj, A. Polanski, M. Staniszewski. Consensus Algorithm for Bi-clustering Analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, pp. 557-570. doi: 10.1007/978-3-031-08754-7_61.

[15] Bioconductor. Available on: https://www.bioconductor.org/developers/package-guidelines/

[16] Bioconductor: Open source software for Bioinformatics. Available on: https://www.bioconductor.org/

[17] M. Carlson. *GO.db:* A set of annotation maps describing the entire Gene Ontology. R package version 3.8.2. 2019.

[18] M. Carlson. *org.Hs.eg.db:* Genome wide annotation for Human. R package version 3.8.2. 2019.

[19] Bioconductor package biomaRt. Available on https://github.com/grimbough/biomaRt.

[20] A. Alexa, J. Rahnenfuhrer. topGO: Enrichment Analysis for Gene Ontology. R package version 2.54.0. 2023.

[21] T. Wu, E. Hu, S. Xu, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation, 2021, vol. 2(3), art. no. 100141. doi: 10.1016/j.xinn.2021.100141.