DOI: https://doi.org/10.15276/hait.07.2024.24
UDC 004.048+004.094

Current state of methods and algorithms for gene expression data clustering and biclustering: A survey

Oleg R. Yarema[1)]
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
Sergii A. Babichev[2,3)]
ORCID: https://orcid.org/0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
[1)] Ivan Franko National University of Lviv, 1, Universytetska Str. Lviv, 79000, Ukraine
[2)] Jan Evangelista Purkyně University in Ústí nad Labem, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic
[3)] Kherson State University, 14b, Shevchenko Street. Sivka-Voynylivska, Ivano-Frankivsk Oblast, 77311, Ukraine

ABSTRACT

The analysis of gene expression data has grown increasingly complex with the expansion of high-throughput techniques like bulk RNA-seq and scRNA-seq. These datasets challenge traditional clustering methods, which often struggle with the high dimensionality, noise, and variability in biological data. Consequently, biclustering methods, which group genes and conditions simultaneously, have gained popularity in bioinformatics. Biclustering is valuable for identifying co-regulated gene subsets under specific conditions, aiding in the exploration of transcriptional modules and gene-disease links. This review examines both traditional clustering and biclustering methods for gene expression analysis, covering applications such as patient stratification, gene network identification, and drug-gene interaction studies. Key biclustering algorithms are discussed, focusing on their strengths and challenges in handling complex profiles. The article highlights significant issues like hyperparameter optimization, scalability, and the need for biologically interpretable results. Emerging trends are also reviewed, such as consensus clustering and distance metrics for high-dimensional data, with attention to the limitations of evaluation metrics. The potential for these methods in diagnostic systems for diseases like cancer and neurodegenerative disorders is also considered. Finally, we outline future directions for enhancing clustering and biclustering algorithms to create a personalized medicine system based on gene expression data.

Keywords: Data mining; gene expression data; clustering; biclustering; decision-making system; ensemble-based methods; alternative voting

***For citation*: Yarema O. R., Babichev S. A. "Current state of methods and algorithms for gene expression data clustering and biclustering: A survey". *Herald of Advanced Information Technology*. 2024; Vol.7 No.4: 347–360. DOI: https://doi.org/10.15276/hait.07.2024.24**

INTRODUCTION

The increasing volume of biological data generated by modern experimental techniques, such as DNA microarrays and RNA sequencing, has led to a growing demand for advanced methods and algorithms for analyzing gene expression data. Among the most critical tasks in this area are clustering and biclustering, which allow for identifying groups of genes with similar expression patterns or co-expressed under specific conditions. These methods hold particular promise in developing diagnostic systems for complex diseases, such as cancer and neurodegenerative disorders, where understanding gene expression profiles can lead to improved prediction, diagnosis, and treatment strategies.

The uniqueness of experimental gene expression data lies in its high dimensionality, noise, and heterogeneity. These characteristics make traditional clustering methods insufficient, as they often fail to capture the intricate structure of the data. In contrast, biclustering approaches provide a more refined analysis by allowing the discovery of gene subsets that are co-regulated under specific subsets of conditions. Despite this potential, several challenges remain unresolved in this domain. Among them are issues related to the scalability of algorithms, the accuracy and interpretability of the results, and the integration of different data types to enhance the robustness of the clustering and biclustering processes.

The primary goal of this survey is to provide a comprehensive analysis of the current state of methods and algorithms for clustering and biclustering gene expression data. We aim to highlight the existing problems in this field and assess the effectiveness of various approaches in addressing these challenges. Ultimately, the review create the conditions to the development of a robust diagnostic system for disease prediction based on

gene expression profiles, paving the way for more accurate and personalized medicine.

1. PROBLEM STATEMENT

The analysis of gene expression data, critical for insights into biological processes and disease mechanisms, faces considerable challenges due to the high dimensionality, noise, and variability of data produced by modern sequencing technologies such as bulk RNA-seq and scRNA-seq. Traditional clustering methods often fall short in addressing these complexities, limiting their utility in identifying co-regulated gene groups or conditions. Consequently, there is an increasing need for advanced clustering and biclustering techniques that can simultaneously analyze genes and conditions to uncover transcriptional modules and gene-disease associations. However, despite their potential, current clustering and biclustering methods encounter obstacles related to algorithm scalability, hyperparameter optimization, interpretability, and the integration of biologically meaningful results. Addressing these challenges is essential for advancing diagnostic tools and personalized medicine based on gene expression data.

2. CURRENT ADVANCES IN CLUSTERING METHODS FOR DATA MINING

In recent years, there has been a significant increase in the number of data mining and machine learning methods [1], particularly in clustering [2], which are widely applied in bioinformatics. Each clustering method has its unique properties, but none of them is universally “best” for all types of data. The primary goal of clustering is to organize a dataset into a smaller number of groups (clusters) so that similar elements are grouped together, while dissimilar ones are placed in different groups. The degree of similarity between elements is typically gauged by their “distance”: the closer the elements are, the higher their similarity. Partition-based clustering techniques, which rely on iterative algorithms, strive to identify the optimal K centers to separate the data into K clusters. These centers can either be centroids, as used in the k-means algorithm, or medoids, as employed in the k-medoids algorithm. The k-means algorithm locates centroids by minimizing the total squared Euclidean distances between each data point and its nearest centroid. Its advantage is low computational complexity, but it is sensitive to outliers and requires prior determination of the number of clusters *K*. Many scRNA-seq analysis methods use k-means. For instance, the SAIC method [3] integrates k-means clustering with ANOVA for cell grouping, followed by the identification of gene signatures. SCUBA [4] divides cells into two groups at each time interval and applies gap statistics to identify bifurcation points. SC3 [5] utilizes k-means for projecting pairwise cell distance matrices and merges clustering results via a consensus function. Both pcaReduce [6] and scVDMC [7] use k-means for initializing their algorithms.

The k-medoids algorithm selects *K* points from the original *N* data points in a way that minimizes the total distance to these medoids. This technique performs well with discrete data that have well-defined cluster centers. However, like k-means, it is sensitive to outliers and requires predefining the number of clusters *K*. RaceID2 [8], designed to detect rare cell types in scRNA-seq data, demonstrated that replacing k-means with k-medoids improves clustering accuracy.

Hierarchical clustering remains the most commonly used method for analyzing gene expression data. It constructs a hierarchical structure among data points, which naturally forms clusters through the tree’s branches. Many scRNA-seq clustering algorithms either utilize hierarchical clustering or incorporate it as a step in their analysis. One of its advantages is that it requires few assumptions about the data's distribution, making it applicable to datasets with diverse shapes. Furthermore, hierarchical clustering effectively represents the relationships between all data points, aiding in the interpretation of clustering results. interpretation of clustering results. There are two primary types of hierarchical clustering: agglomerative and divisive. BackSPIN [9] is a bidirectional clustering method that applies hierarchical clustering across both the gene and cell dimensions. The correlation matrix of gene expression data is progressively divided through an iterative process using SPIN [10] until the predefined separation criteria are satisfied. The cellTree [11] algorithm creates a hierarchical structure for individual cells by generating a minimum spanning tree based on distributions produced by Latent Dirichlet Allocation (LDA). CIDR [12] employs hierarchical clustering on top coordinates obtained via Principal Coordinates Analysis (PcoA) from a dissimilarity matrix created after imputing missing data. ICGS [13] uses hierarchical clustering to group gene expression data, selecting genes based on their expression levels and dynamic range, followed by pairwise correlation analysis. RCA [14] applies hierarchical clustering to a correlation matrix that is constructed from the

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

projections of single-cell profiles onto aggregated scRNA-seq data. SC3 [5] also utilizes agglomerative clustering on a consensus matrix, which is formed by merging the results of multiple k-means clusterings. DendroSplit identifies clusters within a hierarchical tree by performing dynamic splits and merges of branches, using a division index based on the original gene expression data.

Mixture model-based clustering operates on the assumption that the data are samples derived from a mixture of different probability distributions, with each distribution representing a distinct cluster. Formation of a cluster structure is carried out by estimating the likelihood of each sample belonging to a specific distribution. For continuous data, the Gaussian Mixture Model (GMM) is the most popular, whereas the categorical mixture model is preferred for discrete data. These approaches provide the benefit of a robust probabilistic framework, enabling the integration of prior knowledge into the clustering process. Nevertheless, interpreting mixture models necessitates advanced optimization or sampling methods, which are computationally demanding and rely heavily on the assumptions made about the data distribution. Typically, training mixture models is performed using the Expectation-Maximization (EM) algorithm, which alternates between estimating mixture parameters and classification probabilities. Alternatively, sampling and variational approaches are employed for training probabilistic graphical models. The computational complexity of these models is determined by the type of distribution used in the mixture.

BISCUIT [15], as an illustration, utilizes a Hierarchical Dirichlet Mixture Model (HDMM) and includes cell-specific scaling along with the imputation of missing values. This model treats cells as a Gaussian mixture with a Dirichlet distribution for the mixing coefficients, a normal prior for the means, and a Wishart distribution for the covariance matrices, accounting for technical variability between individual cells. BISCUIT is trained using Gibbs sampling. Seurat 1.0 [16] integrates scRNA-seq data with in situ RNA to enable spatial clustering of cells. The integration is achieved through a bimodal mixture model that focuses on a specific group of marker genes, enabling the identification of spatial clusters based on the probability that a scRNA-seq profile belongs to a given cluster. DTWScore [17] selects the most distinctive genes from scRNA-seq time-series data and uses GMM to cluster cells. TSCAN [18] employs GMM to cluster cells and then constructs a minimum spanning tree is employed to identify pseudotemporal progression.

Graph-based clustering models data points as nodes in a graph, with edges representing pairwise similarities between them. This method operates on the assumption that dense communities exist within the graph, which can be visualized as dense subgraphs or spectral components. Although these algorithms are less reliant on assumptions about data distributions, they often require substantial computational resources, which is a significant limitation. Spectral clustering and clique detection are among the most widely used graph-based clustering algorithms. Spectral clustering [19] begins by constructing a similarity matrix and a graph Laplacian using a similarity function, such as the RBF kernel (which must be tuned). The eigenvectors of the Laplacian are calculated, and k-means is used for the clustering process. However, the high computational load involved in calculating all eigenvectors often makes spectral clustering impractical for large datasets. In TCC-based clustering [20], spectral clustering is applied with a similarity matrix based on transcript compatibility and Jensen-Shannon divergence between cells, but only when the number of cell types is predetermined; otherwise, affinity propagation is employed. SIMLR [21] refines cell similarity metrics by introducing rank constraints and graph diffusion, followed by spectral clustering on the latent components.

A clique in graph theory refers to a subgraph where each node is connected to every other node, representing clusters of data points within the graph. Since finding cliques can be computationally expensive, heuristic approaches are often employed. For instance, SNN-Cliq [22] uses clique detection to cluster cells based on scRNA-seq data. Since true cliques are rare in sparse graphs, it instead identifies dense, quasi-cliques that are not fully connected in the shared nearest neighbor (SNN) graph. In single-cell analysis, one of the most widely used graph-based clustering methods is the Louvain algorithm, a community detection approach that provides better scalability than other methods. It uses a greedy strategy to assign nodes to communities and iteratively updates the network to achieve a coarser representation. For instance, SCANPY [23] integrates the Louvain algorithm for large scRNA-seq dataset analysis, and Seurat [16] uses it on an SNN graph to classify cell types.

Density-based clustering identifies clusters as areas in the input space with a high density of data points. Notable examples include DBSCAN and

density peak clustering. DBSCAN [24] forms clusters by choosing a data point as the center of a sphere with radius ε and checking whether the number of points inside the sphere surpasses a given threshold. This operation is performed for every point to expand the clusters. This method is both efficient and versatile for data of arbitrary shapesıo However, DBSCAN is sensitive to parameter tuning, and its performance may degrade if cluster densities are uneven. In scRNA-seq analysis, density-based clustering is frequently used to find outlier cells, as illustrated by GiniClust, and Monocle 2 [25]. Unlike DBSCAN, which uses a density threshold, density peak clustering [26] focuses on the distances between points and assumes that cluster centers are local maxima of density. Monocle 2 [25] uses this technique for cell clustering within the t-SNE-generated space.

Kohonen networks, commonly referred to as self-organizing maps (SOMs) [27], employ competitive learning to perform clustering tasks. This method iteratively updates the cluster center positions for each data point, with adjustments weighted based on the similarity or distance between the point and centers, using stochastic gradient descent. Cluster centers are commonly initialized on predefined structures like grids. SOMs are scalable because stochastic gradient descent does not require all data points to be stored in memory simultaneously. Additionally, these predefined structures can incorporate prior knowledge, allowing for more interpretable relationships between clusters to be established. Despite their advantages, SOMs are quite sensitive to parameter choices, especially the learning rate in weight adjustments. They are commonly utilized for visualizing and clustering scRNA-seq data. Studies [28] have utilized SOMs to create intuitive visualizations, such as 2D heatmaps, where the spatial layout reflects similarities in expression patterns. The SCRAT software package [29] supports the generation of 2D heatmaps that display correlations between genes in single-cell profiles. SOMSC [30] applies SOMs to compress high-dimensional gene expression data into a two-dimensional space, aiding in the detection of transitions between cell states and in ordering cells along a pseudotemporal trajectory.

An enhancement of traditional self-organizing maps (SOMs) is the Self-Organizing Tree Algorithm (SOTA) [31], which allows the formation of a tree structure, providing better representation of hierarchical relationships between clusters. Unlike SOMs, SOTA forms a tree, enabling the natural representation of cluster hierarchies, useful for visualizing and analyzing complex data relationships. SOTA adapts to the data by splitting tree nodes as necessary, offering more flexible and accurate clustering compared to the fixed grid structure of SOMs. Due to its hierarchical structure, SOTA is more efficient with large datasets, as it does not require processing all points simultaneously, which can be problematic for SOMs. The tree structure of SOTA facilitates the interpretation of clusters and their relationships, while SOM provides a less informative flat map. Thus, SOTA offers a more flexible and effective approach to clustering gene expression data compared to traditional SOMs.

Ensemble clustering, or consensus clustering, uses multiple approaches to cluster the same dataset. The results are then combined using a consensus function. This approach accounts for the diversity of data representations and clustering models, making it more robust and effective compared to individual models. Nonetheless, the effectiveness of ensemble clustering is influenced by the quality of the underlying algorithms and the methods used for data transformation. SC3 [5] is an example of a consensus approach designed for clustering scRNA-seq data. It initiates by computing pairwise distance matrices between cells based on Euclidean, Pearson, and Spearman correlations, then applies transformations via PCA and the Laplacian. The six resulting projections are then clustered using the k-means algorithm, and the outcomes are combined into a consensus matrix by employing a similarity-based partitioning strategy [5]. This matrix is subsequently used for hierarchical clustering. Another consensus method, conCluster [32], merges multiple partitions derived from different runs of t-SNE (t-Distributed Stochastic Neighbor Embedding) and k-means, each with varying parameters, and merges them for final clustering using k-means. Therefore, SC3 and conCluster demonstrate the advantages of ensemble clustering by providing higher accuracy and robustness through the combination of different approaches.

The Affinity Propagation clustering algorithm [33] works by exchanging messages between two types of log-likelihoods to determine cluster centers (exemplars). The first type of message, responsibility, indicates how appropriate a data point xk is for representing another point xi compared to other possible candidates. The second type, availability, evaluates how appropriate it is for point x_i to be represented by point x_k, considering other points also represented by x_k. The main advantage of Affinity Propagation is that it does not require the

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

number of clusters to be predefined. However, Affinity Propagation has some limitations, including its high computational cost and sensitivity to outliers. In TCC-based clustering [20], Affinity Propagation is applied for cell clustering when the number of cell types isn't predetermined. Additionally, SIMLR [21] can employ Affinity Propagation on the similarity matrix derived from multiple kernels, without the need for spectral clustering in the latent space.

3. ENSEMBLE CLUSTERING IN GENE EXPRESSION DATA ANALYSIS

When analyzing gene expression data, experimental data can be structured in multiple ways. In the conventional approach, the data is typically arranged as a matrix where the rows signify genes, and the columns represent samples (such as various tissues or conditions). In this structure, each matrix element shows the expression level of a specific gene in a particular sample. This format is common in bulk RNA-seq analysis, which focuses on gene expression across different samples. In the case of single-cell RNA-seq (scRNA-seq), the matrix is transposed, with rows corresponding to cells and columns representing genes. Each element in this matrix reflects the expression level of a specific gene in an individual cell. For cancer research, scRNA-seq can help identify specific genes or gene combinations that are characteristic of cancer cells. This aids not only in diagnosis but also in developing personalized treatment strategies by identifying the most effective therapeutic targets for each patient.

It is important to note that, regardless of the specific approach, the initial data are always represented as a matrix of gene expression values:

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{m1} & \cdots & e_{mn} \end{bmatrix}, \qquad (1)$$

where: e_{ij} is the expression level of the j-th gene corresponding to the i-th sample or cell; n is the number of samples or cells; m is the number of genes.

In the context of bulk RNA-seq or scRNA-seq data analysis, samples are commonly separated into k non-overlapping clusters through a chosen clustering algorithm. This allows for the grouping of relevant sample or cell types within the gene expression matrix based on their characteristics. Accurate clustering of genes enables the identification of significant gene subsets, which can be further utilized in diagnostic models. These models use the identified gene clusters as attributes for classifying samples or cells. The application of ensemble clustering methods at this stage enhances the objectivity of decision-making regarding the nature of the grouping of the studied objects. Fig. 1 illustrates a general scheme of the method based on the application of a cluster ensemble [34]. Also known as consensus clustering or cluster aggregation, ensemble clustering aims to recover the inherent groupings of samples or cells by utilizing labels from different data partitions [35]. The key objective is to integrate several base clustering results into a unified clustering solution, as depicted in Fig. 1. To date, various methods and approaches have been proposed and applied to gene expression data processing, such as cola [36], scEFCS [37], SC3 [5], and SHARP [38]. While these approaches address different scientific challenges and focus on distinct aspects, the fundamental principles and key issues surrounding the generation and integration of numerous partitions or models remain consistent, enhancing the objectivity of cluster structure formation. This approach can also be extended to bulk RNA-seq data analysis, where instead of individual cells, samples are considered, which similarly require reliable and accurate clustering to uncover natural groupings of samples and/or genes.

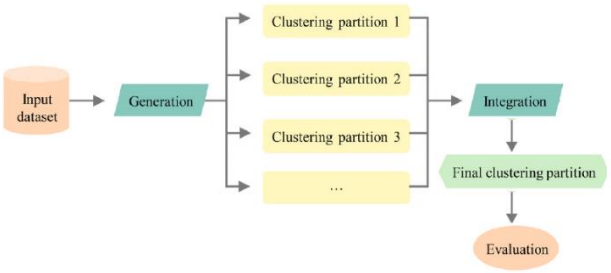

Fig. 1. **General scheme of the ensemble clustering method**
Source: **compiled by the authors**

The development of multiple diverse cluster partitions or models is a key requirement for all ensemble clustering techniques. Ensemble clustering techniques currently applied in single-cell and multi-cell transcriptomics can be broadly divided into three types: gene-oriented methods, cell-oriented methods, and strategies that emphasize various algorithms (Fig. 2) [34]. A typical approach for generating base cluster partitions involves randomly selecting a set number of gene features from the gene expression matrix, forming subsets that capture only part of the original genetic data. Repeating this process multiple times results in a series of subsets, each used for further cluster analysis. To ensure the accuracy of the subsequent analysis, a preprocessed

gene expression matrix, where the dimensionality is reduced to retain the most significant genes, is usually employed. For instance, the cola method [36] starts by performing feature selection to prepare the data, then generates multiple cluster partitions, offering the option to sample either genes or cells depending on user specifications. Random selection of cells, similar to gene sampling, is also a key strategy in gene expression data analysis.

One common approach for creating subsets of gene expression data involves sorting gene features according to a predefined criterion. This allows for obtaining a data subset that retains a subset of the original gene information after dimensionality reduction. Once the dimensionality of the initial data is reduced, the remaining or newly formed features are listed in descending order according to their variability. Due to the fact that each feature impacts the result differently on the final outcome, the features with the highest variability values are typically retained as low-dimensional input data for further analysis. In the case of high-dimensional data, dimensionality reduction helps uncover information hidden within these higher dimensions. However, this process can also result in the loss of critical information by removing part of the structural details. To mitigate this drawback, it is typically recommended to combine several dimensionality reduction methods [39]. Practically, employing different reduction techniques produces multiple data subsets. The resulting clustering's from these subsets are integrated into a final outcome, overcoming the limitations of a single dimensionality reduction technique.

Feature selection is a widely used dimensionality reduction technique that removes genes with lower variability, keeping those with higher variability for cluster analysis. The process starts with calculating measures of variability for each gene, including standard deviation, variance, and coefficient of variation. After arranging the genes in descending order by these indices, the top features are selected to generate data subsets. Feature extraction plays a central role as a dimensionality reduction method in high-dimensional single-cell expression profiles. In contrast to feature selection, classical feature extraction techniques, like PCA, produce "new components" by combining several gene features. The biological significance of these principal components has not been thoroughly explored. However, this does not indicate that the "new principal components" are lacking in value. Researchers commonly select different subsets of principal components to capture all available variability and represent the original dataset in practical applications. In [40], the ANMF-CE method produces multiple base cluster partitions by selecting new dimensions after feature extraction. The study utilizes the Adaptive Total Variation Non-Negative Matrix Factorization (ATV-NMF) algorithm for feature extraction, a method that handles missing values, noise, and arbitrarily shaped clusters.

Random projection (RP), unlike other dimensionality reduction techniques, does not necessitate calculating distances or similarities between cells or "new components", thus reducing execution time and resource consumption while still preserving the variability of information in low-dimensional data with high probability, similar to that in the context of high-dimensional data, SHARP is a notable clustering ensemble system based on the random projection method [38]. It applies random projection repeatedly to the matrix to create multiple low-dimensional datasets, which replace the original data for hierarchical clustering, producing a variety of cluster partitions.

Autoencoders are among the most frequently used neural network models, capable of extracting both linear and non-linear features from the original data. Different types of autoencoders (AEs) are commonly applied to reduce high-dimensional scRNA-seq data into lower dimensions. As an illustration, the scIAE method outlined in [41] separates gene expression profiles into training and testing subsets, applying random projection to each one independently to create several subsets of the original data.

Another approach involves splitting single-cell data into several subsets (subspaces) and analyzing them to generate multi-cluster distributions. It is important to note that these data subsets differ significantly from the primary data matrix. The same genes remain in these submatrices, but the key difference lies in the selection of cells (Fig. 2). In [42], a method based on random cell sampling for ensemble clustering was proposed. The Cola method introduced in [36] relies on repeated random sampling and reclustering of genes or cells, ultimately resulting in a stable clustering outcome. Compared to gene sampling, clustering results from cell sampling subsets tend to be more reliable. Although random sampling allows for the rapid creation of numerous data subsets, it complicates the formation of a subset that fully reflects the original dataset. To achieve stable results and increase the number of sampling iterations, a random stratified

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

sampling strategy can also be applied to the initial single-cell expression data. Unfortunately, these strategies frequently result in heavy computational resource consumption when processing large-scale datasets. [43]. In this study, the authors proposed a new RC approach for cell selection, where a random portion of cells is initially sampled, and then k-means clustering is performed to identify representative cells that encapsulate the entire original dataset [43]. In [44], the authors suggested a similar strategy, but with the difference that the center of each cluster, identified by the k-means method, is considered the representative cell. It's important to recognize that randomly sampling a portion of cells from the original dataset can result in overlapping information. The SHARP algorithm addresses this by dividing the large dataset into equal blocks before sampling. This method optimizes computational resource use, prevents memory overload, and minimizes sampling imbalances. [38].

4. ALGORITHM-ORIENTED METHODS FOR GENE EXPRESSION DATA CLUSTERING

Both gene-based and cell-based approaches, which focus on creating subsets of gene expression data, can result in the loss of some information from the original dataset. In contrast, using methods based on distance metrics and/or clustering algorithms (Fig. 2) allows the preservation of all valuable information contained in the original data. When conducting cluster analysis of gene expression data, two key aspects must be considered: the methods for assessing the distance or similarity between samples or genes, and the algorithm for grouping the relevant data based on these similarity measures.

The effectiveness of applying a partition-based clustering algorithm largely depends on the ability method for assessing the distance or similarity between samples or genes [10]. Numerous approaches are available for measuring the distance between objects, such as Euclidean, Manhattan, Mahalanobis, and Minkowski distances. Similarly, the distance between samples or cells can be assessed through their similarity, where greater similarity corresponds to a smaller distance. This principle is the foundation of several popular similarity measures, such as Pearson and Spearman correlation coefficients, and mutual information scores. By using various distance or similarity metrics, multiple covariance matrices can be created for sample pairs, leading to the generation of several clustering partitions through a specific clustering algorithm (Fig. 2). As different distance or similarity metrics emphasize distinct features of the input data, the final clustering result obtained by combining several partitions from different algorithms tends to be more robust and reliable. Building on this idea, the authors in [5] proposed SC3 (single-cell consensus clustering), a consensus clustering approach for single-cell data analysis that utilizes three common distance and similarity measures: Euclidean distance, Pearson, and Spearman correlation coefficients, ensuring more stable clustering results. Furthermore, the authors in [41] explored four metrics, including those three, and introduced a consensus distance. However, it should be noted that the issue of selecting the most appropriate distance or similarity metric for high-dimensional gene expression data remains unresolved. Euclidean distance is often ineffective for high-dimensional data, and metrics based on mutual information require careful determination of the method for estimating Shannon entropy, which opens the door for further research in this field.

When the clustering process is implemented, a common issue arises in ensuring the appropriate grouping of clustering objects into distinct clusters. The use of different clustering algorithms for gene expression data often leads to inconsistent clustering outcomes, a phenomenon referred to as algorithmic preference in clustering. Applying various algorithms to the same dataset has the potential to generate numerous base clustering partitions, and integrating these results may yield a more reliable clustering outcome. However, this introduces the challenge of optimizing the hyperparameters for each algorithm, which can significantly influence the resulting cluster structures. The scEFCS method, as an example, integrates nine popular clustering algorithms or software tools often used for gene expression data, such as SC3, Monocle, CIDR, pcaReduce, Rphenograph, Seurat, SHARP, SINCERA, and RaceID [37]. In a similar approach, the ECBN framework incorporates four well-known clustering methods or packages for normalized datasets, including CIDR, Seurat, SC3, and t-SNE + k-means [45]. GeoWaVe, developed by Burton and colleagues, utilizes five commonly used clustering algorithms [46], such as FlowSOM [47], PHATE with k-means, SPADE, Phenograph [48], and PARC [49].

As shown in Fig. 3, the alternative voting strategy is currently the most widely adopted approach for obtaining a final result from several base clustering partitions [34]. Due to its simplicity, the majority rule (i.e., voting) provides a stable and representative clustering outcome that reflects the majority of base clustering partitions.

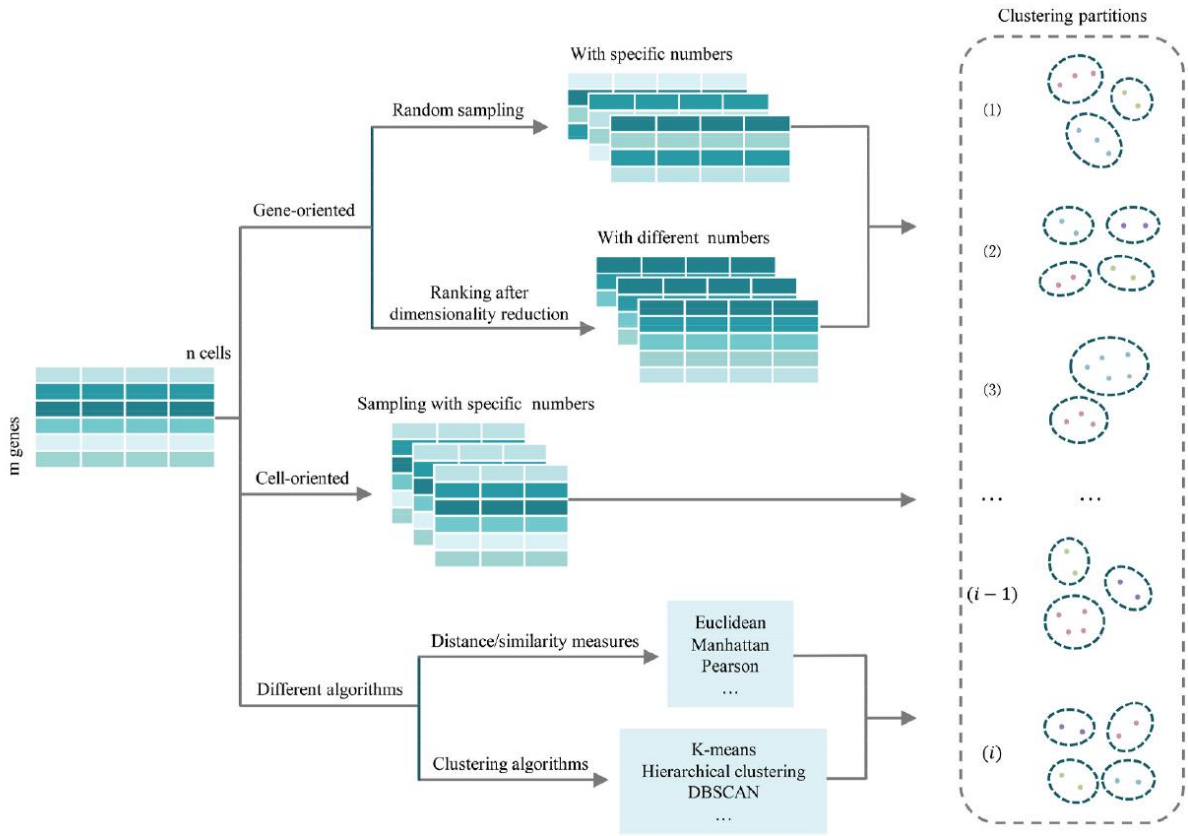

Fig. 2. **Three techniques for generating various cluster partitions**
Source: **compiled by the authors**

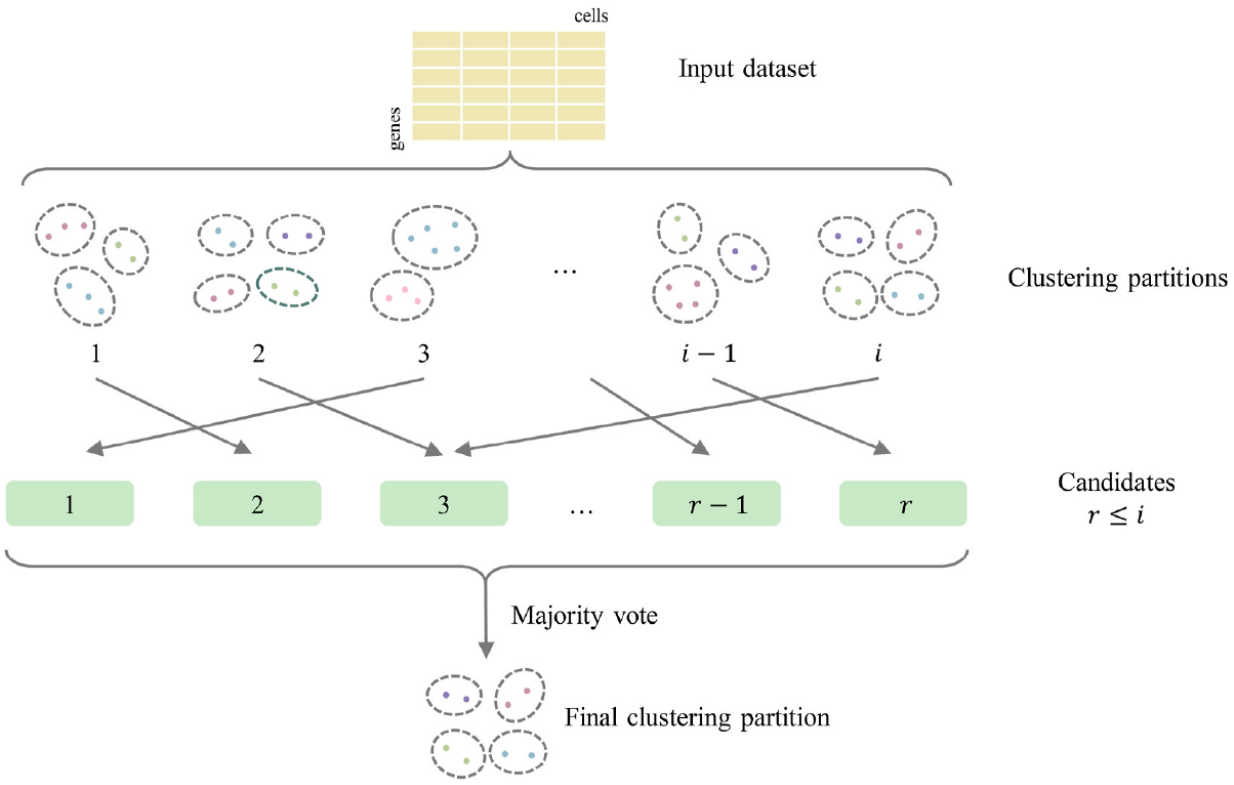

Fig. 3. **Optimal cluster structure selection strategy based on the alternative voting method**
Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

It is important to note that achieving a more accurate clustering result requires a comprehensive input of base clustering partitions, which can lead to increased computational costs. The hypergraph-based approach offers a more adaptable way of representing relationships between data points, which is especially beneficial when integrating multiple base clustering partitions. This approach represents the graph as a diagram illustrating the connections and topological structure between data points. A typical graph is expressed as G (V, E), with V as the vertices and E as the edges. A hypergraph, a more generalized form, allows edges to connect multiple vertices and is denoted H (V, E). Unlike standard graphs, hyperedges can connect more than two vertices. In this case, clustering partitions can be represented as hypergraphs, where cluster labels are transformed into corresponding hyperedges. Each clustering partition p_i is represented using a binary matrix, with the rows corresponding to cells (vertices) and the columns to clusters (hyperedges). The matrix elements v_{jk} indicate the value of the j-th row in the k-th hypergraph.

The following rules govern the assignment of cell labels:

$$v_{ik} = \begin{cases} 1, & if\ the\ i-th\ object\ \in k-th\ cluster \\ 0, & otherwise \end{cases}.$$

The binary matrix assigns the value 1 to elements of a hyperedge when a cell belongs to a particular cluster and 0 when it does not. As a result, each cluster becomes a hyperedge, and the clustering result is illustrated as a hypergraph. The hypergraphs created from different clustering partitions can then be merged into a single large hypergraph to consolidate all partitions. However, one significant downside of the hypergraph strategy is its growing computational complexity as the number of vertices and edges increases. Due to this high computational complexity, applying hypergraph-based clustering to large datasets becomes challenging.

5. METHODS AND MODELS FOR BICLUSTERING GENE EXPRESSION DATA

Biclustering is a data mining technique that groups both rows (observations) and columns (attributes) simultaneously in a data matrix. It enhances traditional clustering methods by exposing more complex relationships between data elements. Fig. 4 visualizes the key differences between clustering and biclustering [50].

Fig. 4 illustrates that clustering methods identify mutually exclusive groups of rows or columns in a data matrix, while biclustering methods uncover data subsets that fulfill criteria of homogeneity and statistical significance (local model). In the figure, orange and blue signify two row clusters (A), two column clusters (B), and two overlapping biclusters (C).

The advantages of bicluster analysis compared to traditional clustering algorithms are as follows: Firstly, biclustering considers similarity between observations (rows) only within a specific subset of attributes (columns), unlike traditional clustering, which considers all attributes when calculating similarity. This makes biclustering particularly useful for biological data analysis, where local patterns exist, such as gene expression data. It allows for the identification of transcriptional modules, which consist of subsets of genes (rows) that correlate within a subset of samples (columns) [51]. Secondly, biclustering permits overlapping groups, meaning that both observations and attributes can belong to multiple groups simultaneously (whereas traditional clustering assigns observations strictly to a single group). This reflects the fact that genes can participate in multiple biological processes simultaneously [52]. Additionally, biclustering provides more flexibility in uncovering complex relationships between observations, making it possible to capture hidden structures and patterns that may not be visible when relying on global models [53]. The strength of biclustering lies in its versatility to detect the combined effects of multiple biological processes active under different conditions, reveal complex biological patterns, and apply methods suited to each research task's needs [54]. Though introduced by Hartigan in 1972 [55], biclustering gained significant attention in biological and biomedical research following the development of the Cheng and Church algorithm in 2000, which was pivotal in gene expression analysis [51]. Today, biclustering is a cutting-edge technique for investigating correlations between gene subsets and experimental conditions, uncovering biological network modules, patient phenotype stratification, and gene-drug relationship analysis [54]. Its applications have also expanded beyond bioinformatics into areas like text mining, recommendation systems, and climatology [56].

Recent research highlights the growing appeal of bicluster-based methods and covers the following directions [6]:

- Algorithmic studies: these focus on analyzing the performance and characteristics of selected biclustering algorithms [52, 53];

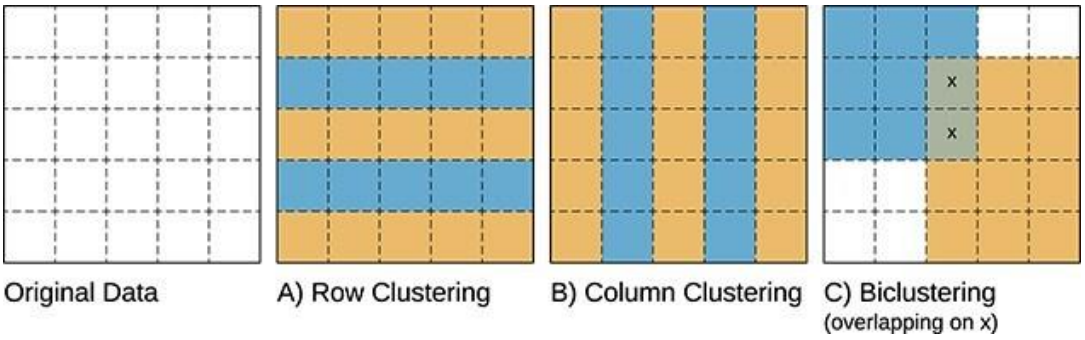

Fig. 4. **The difference between data clustering and biclustering**
Source: **compiled by the authors**

- Comparative studies: these quantitatively compare the effectiveness of different biclustering algorithms [57];
- Evaluation methodologies: these analyze the metrics used by algorithms and in comparative studies to assess performance [58];
- Application studies: these explore the use of biclustering in specific applied fields [54];
- Software studies: these present software tools related to the analysis and application of biclustering methods [54].

However, despite certain advancements in this field, the successful application of bicluster analysis for gene expression data faces several challenges. One of the key issues is the difficulty in accurately defining the boundaries of biclusters, which can lead to the formation of biclusters with heterogeneous structures.

Additionally, the metrics used to evaluate bicluster quality may lack precision or relevance, complicating objective comparison of results. Moreover, the large number of biclusters generated during analysis can hinder result interpretation, particularly when identifying biologically meaningful patterns. Another major obstacle is the sensitivity of algorithms to hyperparameter settings, as even minor adjustments can significantly influence the analysis outcomes, resulting in instability and variability in conclusions. Furthermore, most existing biclustering methods do not account for the more complex multi-level hierarchy of biological processes, limiting their ability to accurately model real biological systems.

Improving bicluster analysis technology in this context may involve enhancing the efficiency of existing algorithms, developing new biclustering algorithms, refining methods for optimizing hyperparameters, improving similarity metrics within biclusters, and increasing the robustness of methods to hyperparameter variations.

CONCLUSIONS

This survey examined the current state of clustering and biclustering algorithms applied to gene expression data, emphasizing their significance in understanding biological processes and disease mechanisms. We highlighted the limitations of traditional clustering methods when faced with the complexity of high-dimensional gene expression datasets. In contrast, biclustering offers a more refined approach, capable of revealing hidden patterns and biological modules by analyzing both genes and experimental conditions simultaneously. The main challenges identified include the optimization of hyperparameters, ensuring the scalability of algorithms, and the need for more interpretable clustering results. Additionally, the effectiveness of biclustering in identifying biologically significant patterns is hindered by the heterogeneity of gene expression data, which complicates the accurate definition of bicluster boundaries. Furthermore, the development of robust evaluation metrics and consensus methods remains critical to improving the reliability of clustering and biclustering outcomes.

The future of clustering and biclustering in gene expression data analysis lies in enhancing existing algorithms and developing new approaches that can effectively model complex biological systems. Priorities include improving the robustness of these methods to handle variability in gene expression data, refining hyperparameter optimization techniques, and incorporating multi-level biological hierarchies into clustering models. These advancements will support the creation of more accurate diagnostic systems and foster progress toward personalized medicine.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

REFERENCES

1. Petegrosso, R., Li, Z. & Kuang, R. “Machine learning and statistical methods for clustering single-cell RNA-sequencing data”. *Briefing in Bioinformatics*. 2020; 21 (4): 1209–1223. DOI: https://doi.org/10.1093/bib/bbz063.

2. Ciaramella, A. & Staiano, A. “On the role of clustering and visualization techniques in gene microarray data”. *Algorithms*. 2019; 12: 123. DOI: https://doi.org/10.3390/a12060123.

3. Yang, L., Liu, J., Lu, Q., et al. “SAIC: an iterative clustering approach for analysis of single cell RNA-seq data”. *BMC Genomics*. 2017; 18 (6): 689. DOI: https://doi.org/10.1186/s12864-017-4019-5.

4. Marco, E., Karp, R. L., Guo, G., et al. “Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape”. *Proc. Natl. Acad. Sci.* 2014; 111 (52): E5643-E5650. DOI: https://doi.org/10.1073/pnas.1408993111.

5. Kiselev, V., Kirschner, K., Schaub, M., et al. “SC3: consensus clustering of single-cell RNA-seq data”. *Nat Methods*. 2017; 14: 483–486. DOI: https://doi.org/10.1038/nmeth.4236.

6. Žurauskienė, J. & Yau, C. “pcaReduce: hierarchical clustering of single cell transcriptional profiles”. *BMC Bioinformatics*. 2016; 17: 140. DOI: https://doi.org/10.1186/s12859-016-0984-y.

7. Zhang, H., Lee, C.-A. A., Li, Z., et al. “A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa”. *PLoS Comput. Biol.* 2018; 14 (4): e1006053. DOI: https://doi.org/10.1371/journal.pcbi.1006053.

8. Grün, D., Muraro, M. J., Boisset, J.-C., et al. “De novo prediction of stem cell identity using single-cell transcriptome data”. *Cell Stem Cell*. 2016; 19 (2): 266–277. DOI: https://doi.org/10.1016/j.stem.2016.05.010.

9. Zeisel, A., Muńoz-Manchado, A. B., Codeluppi, S., et al. “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq”. *Science*. 2015; 347 (6226): 1138–1142. DOI: https://doi.org/10.1126/science.aaa1934.

10. Thrun, M. C. “Distance-based clustering challenges for unbiased benchmarking studies”. *Scientific Reports*. 2021; 11: 18988. DOI: https://doi.org/10.1038/s41598-021-98126-1.

11. Yotsukura, S., Nomura, S., Aburatani, H., et al. “CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data”. *BMC Bioinformatics*. 2016; 17 (1): 363.DOI: https://doi.org/10.1186/s12859-016-1175-6.

12. Lin, P., Troup, M. & Ho, J. W. K. “CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data”. *Genome Biol.* 2017; 18 (1): 59. DOI: https://doi.org/10.1186/s13059-017-1188-0.

13. Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., et al. “Single-cell analysis of mixed-lineage states leading to a binary cell fate choice”. *Nature*. 2016; 537 (7622): 698. DOI: https://doi.org/10.1038/nature19348.

14. Li, H., Courtois, E. T., Sengupta, D., et al. “Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors”. *Nature Genetics*. 2017; 49 (5): 708. DOI: https://doi.org/10.1038/ng.3818.

15. Prabhakaran, S., Azizi, E., Carr, A., et al. “Dirichlet process mixture model for correcting technical variation in single-cell gene expression data”. *In: International Conference on Machine Learning. New York, NY, USA: JMLR.org*. 2016: 1070–1079.

16. Rahul, S., Farrell, J. A., Gennert, D., et al. “Spatial reconstruction of single-cell gene expression data”. *Nature Biotechnology* 2015; 33 (5): 495. DOI: https://doi.org/10.1038/nbt.3192.

17. Wang, Z., Jin, S., Liu, G., et al. “DTWscore: differential expression and cell clustering analysis for time-series single-cell RNA-seq data”. *BMC Bioinformatics*. 2017; 18 (1): 270. DOI: https://doi.org/10.1186/s12859-017-1647-3.

18. Ji, Z. & Ji, H. “TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis”. *Nucleic Acids Res*. 2016; 44 (13): e117. DOI: https://doi.org/10.1093/nar/gkw430.

19. Ng, A. Y., Jordan, M. I. & Weiss, Y. “On spectral clustering: analysis and an algorithm”. *In: Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: MIT Press*. 2002: 849–856.

20. Ntranos, V., Kamath, G. M., Zhang, J. M., et al. “Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts”. *Genome Biology*. 2016; 17 (1): 112. DOI: https://doi.org/10.1101/036863.

21.Wang, B., Zhu, J., Pierson, E., et al. “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning”. *Nature Methods*. 2017; 14 (4): 414. DOI: https://doi.org/10.1038/nmeth.4207.

22. Xu, C. & Su, Z. “Identification of cell types from single-cell transcriptomes using a novel clustering method”. *Bioinformatics*. 2015; 31 (12): 1974-1980. DOI: https://doi.org/10.1093/bioinformatics/btv088.

23. Wolf, A. F., Angerer P. & Fabian J. “SCANPY: Large-scale single-cell gene expression data analysis”. *Genome Biology*. 2018; 19 (1): 15. DOI: https://doi.org/10.1186/s13059-017-1382-0.

24. Ester, M., Kriegel, H.-P., Sander, J., et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. *In KDD, Portland, Oregon: AAAI Press*. 1996; 96: 226–231.

25. Qiu, X., Mao, Q., Tang, Y., et al. “Reversed graph embedding resolves complex single-cell trajectories”. *Nature Methods*. 2017; 14 (10): 979. DOI: https://doi.org/10.1038/nmeth.4402.

26. Rodriguez, A. & Laio, A. “Clustering by fast search and find of density peaks”. *Science*. 2014; 344 (6191): 1492–1496. DOI: https://doi.org/10.1126/science.1242072.

27. Kohonen, T. “The self-organizing map.” *Proc IEEE*. 1990; 8 (9): 1464–1480. DOI: https://doi.org/10.1109/5.58325.

28. Lv, D., Wang, X., Dong, J., et al. “Systematic characterization of lncRNAs’ cell-to-cell expression heterogeneity in glioblastoma cells”. *Oncotarget*. 2016; 7 (14): 18403. DOI: https://doi.org/10.18632/oncotarget.7580.

29. Camp, J. G., Sekine, K., Gerber, T., et al. “Multilineage communication regulates human liver bud development from pluripotency”. *Nature*. 2017; 546 (7659): 533. DOI: https://doi.org/10.1038/nature22796.

30. Peng, T., Nie, Q. “SOMSC: self-organization-map for high-dimensional single-cell data of cellular states and their transitions”. *bioRxiv*. 2017. p. 124693. DOI: https://doi.org/10.1101/124693.

31. Herrero, J., Valencia, A. & Dopazo, J. “A hierarchical unsupervised growing neural network for clustering gene expression patterns”. *Bioinformatic*s. 2005; 17: 126-136. DOI: https://doi.org/10.1093/bioinformatics/17.2.126.

32. Gan, Y., Li, N., Zou, G., et al. “Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method”. *BMC Medical Genomics*. 2018; 11 (6): 117. DOI: https://doi.org/10.1186/s12920-018-0433-z.

33. Afzal, M., Manzoor, I. & Kuipers, O. P. “A fast and reliable pipeline for bacterial transcriptome analysis case study: Serine-dependent gene regulation in Streptococcus pneumonia”. *Journal of Visualized Experiments*. 2015; 98: e52649. DOI: https://doi.org/10.3791/52649.

34. Nie, X., Qin, D., Zhou, X., et al. “Clustering ensemble in scRNA-seq data analysis: Methods, applications and challenges”. *Computers in Biology and Medicine*. 2023; 159: 106939. DOI: https://doi.org/10.1016/j.compbiomed.2023.106939.

35. Golalipour, K., Akbari, E., Hamidi, S.S., Lee, M. & Enayatifar, R. “From clustering to clustering ensemble selection: A review”. *Engineering Applications of Artificial Intelligence*. 2021; 104: 104388. DOI: https://doi.org/10.1016/j.engappai.2021.104388.

36. Gu, Z., Schlesner, M., Hübschmann, D. “cola: an R/Bioconductor package for consensus partitioning through a general framework”. *Nucleic Acids Research*. 2021; 49 (3): e15. DOI: https://doi.org/10.1093/nar/gkaa1146.

37. Bian, C., Wang, X., Su, Y., et al. “scEFSC: Accurate single-cell RNA-seq data analysis via ensemble consensus clustering based on multiple feature selections”. *Computational and Structural Biotechnology Journal*. 2022; 20: 2181-2195. DOI: https://doi.org/10.1016/j.csbj.2022.04.023.

38. Wan, S., Kim, J. & Won, K. J. “SharP: Hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection”. *Genome Research*. 2020; 30 (2): 205-213. DOI: https://doi.org/10.1101/gr.254557.119.

39. Ronan, T., Qi, Z. & Naegle, K. M. “Avoiding common pitfalls when clustering biological data”. *Science Signaling*. 2016; 9 (432): re6. DOI: https://doi.org/10.1126/scisignal.aad1932.

40. Zhu, Y.-L., Gao, Y.-L., Liu, J.-X., Zhu, R. & Kong, X.-Z. “Ensemble adaptive total variation graph regularized NMF for single-cell RNA-seq data analysis”. *Current Bioinformatics*. 2021; 16: 1014-1023. DOI: https://doi.org/10.2174/1574893616666210528164302.

41. Yin, Q., Wang, Y., Guan, J. & Ji, G. “scIAE: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data”. *Briefings Bioinformatics*. 2022; 23: bbab508. DOI: https://doi.org/10.1093/bib/bbab508.

42. Risso, D., Purvis, L., Fletcher, R. B., et al. “Purdom, cluster Experiment, RSEC, A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets”. *PLoS Computational Biology*. 2018; 14: e1006378. DOI: https://doi.org/10.1371/journal.pcbi.1006378.

43. Ringeling, F. R. & Canzar, S. “Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data”. *Genome Results*. 2021; 31: 677–688. DOI: https://doi.org/10.1101/gr.267906.120.

44. Hu, L., Zhou, J., Qiu, Y. & Li, X. “An ultra-scalable ensemble clustering method for cell type recognition based on scRNA-seq data of Alzheimer’s disease”. *In: Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers*. 2022; 275–280. DOI: https://doi.org/10.1145/3544109.3544160.

45. Zhang, D. & Zhu, Y. “ECBN: Ensemble Clustering Based on Bayesian Network Inference for Single-Cell RNA-Seq Data”. *2020 39th Chinese Control Conference (CCC), IEEE*. 2020. p. 5884–5888. DOI: https://doi.org/10.23919/CCC50068.2020.9188589.

46. Burton, R. J., Cuff, S. M., Morgan, M. P., Artemiou, A. & Eberl, M. “GeoWaVe: Geometric Median Clustering with Weighted Voting for Ensemble Clustering of Cytometry Data”. *bioRxiv*. 2022. DOI: https://doi.org/10.1093/bioinformatics/btac751.

47. Quintelier, K., Couckuyt, A., Emmaneel, A., et al. “Analyzing high-dimensional cytometry data using FlowSOM”. *Nat. Protoc.* 2021; 16: 3775–3801. DOI: https://doi.org/10.1038/s41596-021-00550-0.

48. Levine, J. H., Simonds, E. F., Bendall, S. C., et al. “Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis”. *Cell*. 2015; 162: 184–197. DOI: https://doi.org/10.1016/j.cell.2015.05.047.

49. Stassen, S. V., Siu, D. M. D., Lee, K. C. M., et al. “PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells”. *Bioinformatics*. 2020; 36: 2778–2786. DOI: https://doi.org/10.1093/bioinformatics/btaa042.

50. Castanho, E. N., Aidos, H. & Madeira, S. C. “Biclustering data analysis: a comprehensive survey”. *Briefings in Bioinformatics.* 2024; 25 (4): bbae342. DOI: https://doi.org/10.1093/bib/bbae342.

51. Cheng, Y. & Church, G. M. “Biclustering of expression data. Proceedings”. *International Conference on Intelligent Systems for Molecular Biology*. 2000; 8: 93–103.

52. Henriques, R. & Madeira, S. C. “Biclustering with flexible plaid models to unravel interactions between biological processes”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2015; 12: 738–752. DOI: https://doi.org/10.1109/TCBB.2014.2388206.

53.José-García, A., Jacques, J., Sobanski, V., et al. “Biclustering algorithms based on metaheuristics: A Review. Metaheuristics for machine learning”. *Computational Intelligence Methods and Applications. Springer, Singapore*. 2023. p. 39–71. DOI: https://doi.org/10.48550/arXiv.2203.16241.

54. Xie, J., Ma, A., Fennell, A., et al. “It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data”. *Brief Bioinformatics*. 2019; 20: 1450–1465. DOI: https://doi.org/10.1093/bib/bby014.

55. Hartigan, J. A. “Direct clustering of a data matrix”. *Journal of the American Statistical Association*. 1972; 67 (337): 123–129. DOI: https://doi.org/10.1080/01621459.1972.10481214.

56. Madeira, S. C. & Oliveira, A. L. “Biclustering algorithms for biological data analysis: a survey”. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2004; 1: 24–45. DOI: https://doi.org/10.1109/TCBB.2004.2.

57. Castanho, E. N., Aidos, H. & Madeira, S. C. “Biclustering fMRI time series: a comparative study”. *BMC Bioinformatics*. 2022; 23: 192. DOI: https://doi.org/10.1186/s12859-022-04733-8.

58. Noronha, M. D. M., Henriques, R., Madeira, S. C., et al. “Impact of metrics on biclustering solution and quality: a review”. *Pattern Recognition*. 2022; 127: 108612. DOI: https://doi.org/10.1016/j.patcog.2022.108612.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 16.09.2024
Received after revision 15.11.2024
Accepted 21.11.2024

DOI: https://doi.org/10.15276/hait.07.2024.24
УДК 004.048+004.094

Сучасний стан методів і алгоритмів кластеризації та бікластеризації для аналізу даних експресії генів

Ярема Олег Романович[1)]
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
Бабічев Сергій Анатолійович[2,3)]
ORCID: https://orcid.org/0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
[1)] Львівський національний університет імені Івана Франка, вул. Університетська, 1. Львів, 79000, Україна
[2)] Університет Яна Євангелісти Пуркіне в Усті-над-Лабем, Pasteurova 3632/15, 400 96 Усті-над-Лабем Чеська Республіка
[3)] Херсонський державний університет, вул. Шевченка, 14б. Сівка-Войнилівська, Івано-Франківська обл. 77311, Україна

АНОТАЦІЯ

Аналіз даних експресії генів стає дедалі складнішим через розширення високопродуктивних технологій, таких як bulk RNA-seq та одноядерне секвенування РНК (scRNA-seq). Ці набори даних створюють значні виклики для традиційних методів кластеризації, які часто не здатні справлятися з високою вимірністю, шумом та варіабельністю, властивими біологічним даним. Як результат, у біоінформатиці набувають популярності методи бікластеризації, що дозволяють одночасно групувати гени та умови. Бікластеризація є корисною для ідентифікації підмножин співрегульованих генів за певних умов, сприяючи дослідженню транскрипційних модулів та зв'язків між генами та хворобами. Цей огляд охоплює як традиційні методи кластеризації, так і методи бікластеризації для аналізу експресії генів, розглядаючи їх застосування для стратифікації пацієнтів, ідентифікації генних мереж та дослідження взаємодії між генами та ліками. Обговорено ключові алгоритми бікластеризації з акцентом на їхні сильні сторони та виклики у роботі зі складними профілями. Стаття висвітлює важливі питання, такі як оптимізація гіперпараметрів, масштабованість та необхідність біологічно інтерпретованих результатів. Розглянуто новітні тенденції, такі як консенсусна кластеризація та метрики відстані для високовимірних даних, а також обмеження поточних метрик оцінки. Розглядається потенціал цих методів у діагностичних системах для таких захворювань, як рак та нейродегенеративні розлади. Нарешті, ми окреслюємо перспективні напрями для вдосконалення алгоритмів кластеризації та бікластеризації з метою створення системи персоналізованої медицини на основі даних експресії генів..

Ключові слова: інтелектуальний аналіз даних; дані експресії генів; кластеризація; бікластеризація; система прийняття рішень; методи на основі ансамблів; альтернативне голосування; персоналізована медицина

ABOUT THE AUTHORS

Oleg R. Yarema - Candidate of Engineering Sciences, Associate Professor, Department of Digital economics and Business Analytics, Ivan Franko National University of Lviv, 1, Universytetska St. Lviv 79000, Ukraine
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
***Research field*:** Deep Learning; data mining; gene expression data processing, hybrid models, IoT; clustering; development of IT technologies

Ярема Олег Романович - кандидат економічних наук, доцент кафедри Цифрової економіки та бізнес-аналітики. Львівський національний університет імені Івана Франка, вул. Університетська, 1. Львів, 79000, Україна

Sergii A. Babichev - Doctor of Engineering Science, Professor, Department of Informatics, Jan Evangelista Purkyně University in Ústí nad Labem, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic
Professor of the Department of Physics, Kherson State University, 14b Shevchenko Street, Sivka-Voynylivska, Ivano-Frankivsk Oblast, 77311, Ukraine
ORCID: https://orcid.org/ 0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
***Research field*:** Deep Learning; data mining; gene expression data processing, clustering, classification, hybrid models; Internet of things

Бабічев Сергій Анатолійович - д-р техніч. наук, професор, професор кафедри Інформатики університету Яна Евангеліста Пуркінє в Усті на Лабі, Чехія. Професор кафедри фізики Херсонського державного університету. Херсон, Україна