Global NEST Journal, Vol 26, No 9, 06170

Exploring Gene Expression Biclustering with Integrated PSO-SA-Fuzzy Logic Methodology and its Application to Air Pollution Analysis

S. Deepajothi[1]*, Umamaheswari. M.[2], A. Viswanathan[3] and R. Juliana[4]

[1]Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur-603 203, Tamil Nadu, India

[2]Assistant Professor, School of CSE, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, India

[3]Associate Professor., School of CSE, Vellore Instituteof Technology, Vellore-632014, Tamil Nadu, India

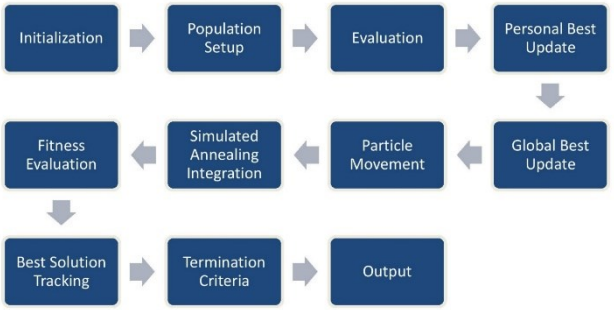
[4]Professor, Department of Information Technology, Loyola ICAM College of Engineering and Technology. Nungambakkam, Chennai-600034, Tamil Nadu, India

Received: 13/05/2024, Accepted: 23/06/2024, Available online: 08/10/2024

*to whom all correspondence should be addressed: e-mail: jothideepa53@gmail.com

https://doi.org/10.30955/gnj.06170

Graphical abstract



Abstract

Gene expression data analysis is crucial for understanding complex biological mechanisms, yet current biclustering techniques struggle with noise, unpredictability, and intrinsic uncertainty. This study proposes an innovative biclustering method combining particle swarm optimization (PSO), simulated annealing (SA), and fuzzy logic to improve gene expression analysis. By integrating natural language processing (NLP) semantic similarity into the PSO framework, the method enhances the capture of intricate gene interactions. Additionally, environmental factors, particularly air pollution, are incorporated to explore their impact on gene expression patterns. The approach leverages the complementary strengths of PSO's exploration capabilities, SA's exploitation efficiency, and fuzzy logic's ability to handle data ambiguity. Comparative assessments on benchmark datasets reveal that this integrated strategy significantly outperforms individual methods in accuracy and resilience. The results demonstrate the PSO-SA-Fuzzy Logic method's superior capacity to detect nuanced and context-dependent gene expression patterns, offering a robust solution to the limitations of existing biclustering techniques. The method not only improves precision and computational efficiency but also enhances the detection of significant gene expression patterns under varying conditions, including environmental stressors. This advancement represents a notable contribution to computational biology, providing a more effective tool for gene expression data analysis.

Keywords: Gene expression biclustering, Particle Swarm Optimization, Simulated Annealing, Fuzzy Logic, Air Pollution

1. Introduction

The discovery of coherent gene subsets under specified conditions is crucial in gene expression analysis, as it plays a critical role in enhancing our understanding of complex biological processes [Orzechowski *et al.* 2019]. The utilization of biclustering as a methodology has arisen as a potentially fruitful approach for elucidating intricate patterns within gene expression data [Jeynes *et al.* 2023 and Ovens *et al.* 2021]. Although biological systems hold great potential, the existing approaches encounter difficulties in effectively collecting complex patterns due to the inherent noise and variability within these systems [Wang and Gao 2019]. The comprehension of the activation or repression of genes under various circumstances is essential for elucidating the complex mechanisms that govern biological systems [Ovens *et al.* 2020]. This analytical undertaking entails examining the expression of genes by means of messenger RNA production, thereby facilitating biological processes [Tercan and Acar 2019]. Air pollution poses a severe threat to public health and the environment, with detrimental effects on respiratory health, cardiovascular diseases, and the ecosystem. Understanding the complex interplay between air pollutants and biological systems is essential for devising effective mitigation strategies. Gene expression analysis offers valuable insights into the molecular mechanisms underlying the response to air

pollution exposure. However, existing methodologies encounter challenges in capturing the dynamic relationship between gene expression patterns and air quality parameters due to the complexity and variability of environmental stressors.

The measurement of gene expression plays a crucial role in understanding the molecular characteristics that drive many biological occurrences, ranging from developmental processes to pathways associated with diseases [Qureshi 2021]. Through the analysis of gene expression, scientists are able to decipher how cells respond to environmental stimuli, find biomarkers associated with diseases, and discover prospective targets for therapeutic interventions [Zapirain *et al.* 2020]. In recent years, there has been a remarkable advancement in high-throughput sequencing and microarray technologies, which have made it possible to assess hundreds of genes simultaneously [Marcos-Zambrano *et al.* 2021]. The recent progress in gene expression analysis has propelled it into the realm of big data, presenting new possibilities for understanding biological complexity on an unprecedented level [Chowdhury *et al.* 2019]. Nevertheless, the abundance of gene expression data presents several difficulties [Ovens *et al.* 2020]. Noise, unpredictability, and the enormous amount of information available all pose difficulties in understanding expression patterns effectively [Mahrishi *et al.* 2020]. The utilization of robust procedures is crucial in order to identify significant signals within the intricate nature of gene expression research. Consequently, innovative computational strategies have been investigated [Jo *et al.* 2021].

The examination of gene expression patterns offers valuable insights into the fundamental molecular mechanisms that regulate cellular functions [Ceglia *et al.* 2022]. Biclustering methods are a useful way to figure out how genes work in different situations, which helps us understand more about complicated biological networks [Sukumar *et al.* 2019]. Nevertheless, the current approaches have certain limitations when confronted with the complexities of gene expression datasets in real-world scenarios [Livochka *et al.* 2023].

The noise, variability, and uncertainty that are naturally present in biological systems make it harder to find gene subsets that show consistent expression patterns. This causes big problems. Existing biclustering techniques frequently struggle to effectively address these difficulties, thereby requiring a more sophisticated strategy.

The challenge at hand pertains to the development of a biclustering approach that effectively tackles the difficulties arising from noise and variability while also including optimization techniques to improve both accuracy and robustness. The objective of this study is to address the existing research gap by examining the possible combined effects of particle swarm optimization, simulated annealing, and fuzzy logic in the domain of gene expression biclustering. The proposed method is novel because it uses NLP semantic similarity in conjunction with fuzzy logic, SA, and PSO to improve gene expression biclustering. The capturing of intricate gene connections and semantic links within the data is improved by this hybrid technique. Furthermore, using data on air pollution to evaluate the effects of the environment on gene expression is a novel multidisciplinary application that offers further understanding of the ways in which external stressors affect biological systems.

The contribution of this study to create a cohesive biclustering framework that effectively utilizes the advantages of particle swarm optimization, simulated annealing, and fuzzy logic. The primary objective is to enhance the precision and resilience of recognizing gene expression patterns in specific circumstances through the utilization of natural language processing (NLP) integrated with semantic similarity. Furthermore, the study aims to perform a comparative analysis in order to assess the effectiveness of the proposed strategy in comparison to existing methods.

The research presented in this study is distinguished by the incorporation of particle swarm optimization, simulated annealing, and fuzzy logic. This integration provides a comprehensive approach to address the existing issues encountered by gene expression biclustering approaches. This paper introduces a revolutionary methodology that enhances accuracy and resilience in the field of computational biology. The approach involves incorporating semantic similarity into particle swarm optimization (PSO), resulting in a useful breakthrough.

2. Related Works

The method in [Ramkumar *et al.* 2022] proposes a novel healthcare biclustering model in the field of healthcare biclustering. This model aims to address the issues related to clustering gene expression data. The objective of this study is to utilize fuzzy C-means (FCM) clustering to discern particular gene activities within different contexts, thereby reducing the redundancy of general gene information components. The evaluation demonstrates the superiority of FCM in terms of average match score and reduced runtime when compared to existing approaches such as PSO-SA and fuzzy logic healthcare biclustering methods. Despite the significant progress made in healthcare biclustering models, there is a discernible research void in comprehending the extent to which the proposed FCM clustering method may be applied to various healthcare datasets. Additional investigation is required to evaluate its efficacy across a wider range of gene expression profiles within healthcare settings.

In order to improve the accuracy of biclustering outcomes, [Chu *et al.* 2023] introduces a preprocessing technique called mean-standard deviation (MSD). This method addresses a crucial element of biclustering algorithms. In order to address the issue of potential noise that may arise during the translation of gene expression data into binary matrices, the Minimum Spanning Degree (MSD) method seeks to minimize the loss of information. Furthermore, the present study offers a unique technique called Weight Adjacency Difference Matrix Binary Biclustering (W-AMBB), which demonstrates efficacy in handling datasets that contain overlapping biclusters. The experimental

findings underscore the robustness of W-AMBB, particularly when used with synthetic datasets, and its biological relevance when applied to real datasets. The efficacy of the MSD preprocessing method in improving biclustering accuracy has shown promise. However, there is a lack of research that examines its usefulness across different types of gene expression datasets. Potential areas for further research could involve examining the resilience of mean squared deviation (MSD) across various datasets that possess different properties and degrees of noise.

The research by [Yelugam *et al.* 2023] looks into how adaptive resonance theory (ART)-based biclustering techniques, specifically BARTMAP and TopoART, can be combined to create a new method called TopoBARTMAP. This will allow biclustering to be used in more situations. This study utilizes topological learning techniques to detect interconnected regions within datasets, hence improving the effectiveness of biclustering and module extraction processes. The results show that TopoBARTMAP is statistically significantly better than previous (bi)clustering techniques at finding different types of biclusters. This is shown by the fact that it was tested on both real-world cancer datasets and synthetic datasets. Nevertheless, there is a lack of research that investigates the ability of TopoBARTMAP to handle larger datasets and various biological scenarios, resulting in a research gap. Examining the efficacy of the aforementioned approach on datasets beyond the realm of cancer-related research would yield a more holistic comprehension of its potential utility.

The article in [Chu *et al.* 2022] explores the challenges biclustering methods face when handling binary data matrices. The study shows the Adjacency Difference Matrix Binary Biclustering (AMBB) algorithm, which tries to find a good balance between how quickly it works and how well it works with binary data. The AMBB algorithm generates an adjacency matrix by utilizing adjacency difference values, which aids in the clustering of genes that have similar reactions across various environments. The experimental findings underscore the considerable practicality of the AMBB approach when used with both synthetic and real datasets. The AMBB algorithm has been developed to tackle the difficulties associated with processing binary data. However, there is a research gap in assessing its scalability when used with gene expression datasets with high dimensions. Additional research is necessary to evaluate the efficacy and effectiveness of the method on extensive datasets that are frequently encountered in the field of genomics research.

The authors in [Liu *et al.* 2023] describe ARBic, a biclustering algorithm that was created to efficiently find important biclusters of different shapes in large gene expression datasets. The objective of ARBic is to achieve a balance between broader and narrower bicluster identification by integrating column-based and row-based techniques. The results of comparative studies conducted on both simulated and real datasets demonstrate that ARBic exhibits improved performance when compared to existing tools. ARBic achieves higher recovery, relevance, and F1 scores, indicating its effectiveness in data analysis. While ARBic has shown excellent performance in detecting biclusters with diverse forms, there is a lack of study on its sensitivity to varying amounts of noise and the characteristics of datasets. Additional investigation is required in order to assess the resilience of ARBic in different datasets, taking into account variations in noise levels and the morphologies of biclusters.

The study in [Adhikary and Acharyya 2022] introduces an enhanced iteration of the Grey Wolf Optimizer, known as the Randomized Move Grey Wolf Optimizer (RM-GWO). This variant is specifically employed for the purpose of detecting biclusters within the Parkinson disease dataset. This study represents a groundbreaking endeavor in the use of biclustering techniques on a dataset that is relevant to diseases. The primary objective of this research is to detect and analyze shifting and scaling pattern-based biclusters. The effectiveness of the suggested method is demonstrated through the use of benchmark functions and statistical testing. The utilization of the RM-GWO algorithm for the purpose of identifying biclusters inside the Parkinson disease dataset is a groundbreaking endeavor. Nevertheless, there is a lack of research that investigates the applicability of RM-GWO to different datasets connected to diseases and comprehends its efficacy when compared to other optimization methods.

The authors in [Charfaoui *et al.* 2023] present a unique methodology for biclustering that uses multi-objective differential evolution (DE) as a means of gene group discovery. The Biclustering Binary Differential Evolution (BBDE) algorithm incorporates a novel mutation operator to improve the detection of biclusters that exhibit both high quality and cohesiveness. The results of these tests demonstrate that the algorithms surpass existing methods that are considered to be at the forefront of the field. This showcases the efficiency of the algorithms in accurately detecting important biclusters across a range of situations and biological datasets. Although multi-objective differential evolution (DE) and BBDE have demonstrated their efficiency, there exists a research gap pertaining to the evaluation of their performance on datasets characterized by varied levels of complexity and noise. Subsequent investigations may delve into the examination of the adaptability and resilience of these algorithms when applied to a more extensive array of biological datasets.

Gupta S et al [Preprint] proposes Multi-Edge-IoT, a full-stack system architecture designed to maximize resources for resource-constrained Internet of Things (IoT) devices. It makes use of clever strategies to improve scalability and energy efficiency among heterogeneous IoT devices by forming a clustered environment and using a routing approach known as Search and Rescue Optimization to connect nodes. The reviewed studies offer valuable insights into the relationship between air pollution exposure and health outcomes. Fisher and Oleksiak [2007] found complex gene expression patterns among populations exposed to pollution, indicating both convergence and divergence in responses [Baskar and Rajaram 2022]. Amnuaylojaroen *et al.* [2024] highlighted the potential association between early-life air pollution

exposure and autism spectrum disorders in children, underscoring the importance of neurodevelopmental research. Block *et al.* [2012] emphasized the detrimental effects of air pollution on neurological health, while Al-Kindi *et al.* [2020] discussed the significant role of air pollution in cardiovascular disease, calling for comprehensive interventions. However, limitations exist, including data heterogeneity, potential confounding factors, limited causality assessment, spatial and temporal variability, and challenges in translating research findings into policy and practice. Addressing these limitations is crucial for advancing our understanding of air pollution's health effects and implementing effective mitigation strategies. Andika, H.K et al [2022] suggests using THD-Tricluster approach, a type of triclustering analysis, to analyze gene expression data for tuberculosis (TB) diagnosis and prognosis. Through the examination of three-dimensional datasets that include observations, attributes, and contexts, this approach seeks to detect triclusters that indicate patterns of shifting and scaling in gene expression over time. These triclusters aid in early illness identification and prognosis by offering insightful information about the dynamic changes in gene expression linked to tuberculosis infection (**Table 1**).

Table 1. Summary

Reference	Biclustering Method	Algorithm Used	Performance Metrics	Results
Ramkumar *et al.* (2022)	FCM Biclustering	PSO-SA, Fuzzy Logic	Average Match Score, Run Time	FCM outperforms PSO-SA and fuzzy logic methods
Chu *et al.* (2023)	W-AMBB Biclustering	Mean-Standard Deviation (MSD) Preprocessing	Robustness, GO Enrichment Analysis	W-AMBB is significantly more robust
Yelugam *et al.* (2023)	TopoBARTMAP Biclustering	Adaptive Resonance Theory (ART), Topological ART	Improvement over other (bi)clustering methods on real datasets	Significant improvement on ordered and shuffled data
Chu *et al.* (2022)	AMBB Biclustering	Adjacency Difference Matrix (ADM)	High Practicability	Practicality demonstrated on synthetic and real datasets
Liu *et al.* (2023)	ARBic Biclustering	Column-Based and Row-Based Strategies	Recovery, Relevance, F1 Score	At least 29% higher scores than the best existing tool
Adhikary and Acharyya (2022)	RM-GWO Biclustering	Randomized Move Grey Wolf Optimizer	Identification of Biclusters in Parkinson Disease Dataset	Improved bicluster identification compared to others
Charfaoui *et al.* (2023)	Multi-objective DE, BBDE	Multi-objective Differential Evolution, BBDE	Biological Relevance, Noise Resistance, Overlap Resistance	Outperformed state-of-the-art algorithms

Significant developments in biclustering techniques are highlighted in the study, including the application of fuzzy C-means clustering, mean-standard deviation preprocessing, and approaches based on adaptive resonance theory. Each approach tackles a different set of issues related to the analysis of gene expression data, such as managing noise, finding biclusters that overlap, and increasing computing effectiveness. In illuminating the complex interactions between gene expression and environmental stressors, the review also emphasizes the significance of context-specific applications, such as air pollution exposure and healthcare statistics. More thorough and flexible biclustering approaches are therefore required, as there are still research gaps in assessing these approaches across a range of datasets and circumstances.

3. Methods

The method described in this study utilizes the combined capabilities of PSO, SA, and fuzzy logic in order to improve the process of gene expression biclustering. The objective of integrating PSO with SA is to enhance the efficiency of the biclustering procedure by effectively managing the trade-off between exploration and exploitation. PSO emulates the collective behavior of particles, enabling the systematic examination of the solution space. On the other hand, SA incorporates a stochastic acceptance of poor solutions, thereby improving the exploitation process. Fuzzy logic is utilized as a means to effectively handle the inherent uncertainties and imprecision that are present in gene expression data. Fuzzy logic plays a significant role in enhancing the modeling of gene expression patterns by accommodating the representation of unclear information.

Our proposed method integrates gene expression analysis with air quality monitoring data to elucidate the impact of air pollution on biological systems. By leveraging advanced computational techniques such as machine learning and statistical modeling, we aim to identify biomarkers and pathways associated with air pollution exposure. Additionally, the study incorporates spatial and temporal analysis of air quality data to assess the variability of gene expression patterns in different environmental contexts. This interdisciplinary approach facilitates a comprehensive understanding of the biological responses to air pollution and informs targeted interventions for mitigating its adverse effects on human health and the environment.

3.1. PSO-SA

The PSO-SA methodology entails the integration of two optimization algorithms, one drawing inspiration from social behavior and the other emulating a probabilistic search process. The integration of PSO and SA in PSO-SA results in an improved optimization process. The PSO algorithm is a nature-inspired optimization technique in

which individuals, represented as particles, traverse the solution space in search of the optimal solution. This phenomenon is shaped by the amalgamation of individual and community knowledge. It can be likened to a social network, wherein particles exchange information in order to collectively enhance their comprehension of the solution space. Simulated annealing, however, replicates the annealing process observed in metallurgy. This feature enables the adoption of suboptimal solutions based on probability, hence introducing a stochastic component. The inherent stochasticity of this method allows for a more extensive exploration of viable solutions, hence potentially circumventing the issue of getting trapped in local optima. The PSO-SA algorithm integrates the global exploration capabilities of PSO with the local refining elements of SA. The objective of this partnership is to attain a more equitable and efficient optimization process by capitalizing on the respective advantages of both algorithms. This will result in an improved exploration of optimal solutions within intricate problem domains.

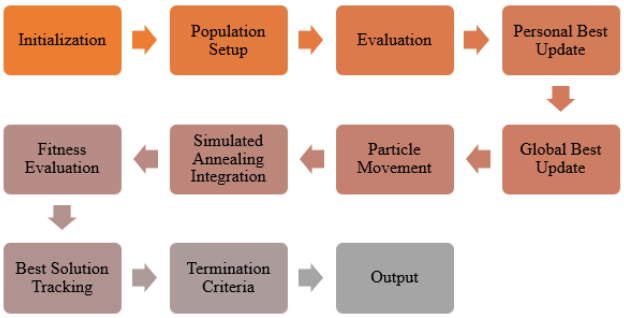


Figure 1. Proposed Framework

The methodology involves collecting comprehensive environmental data using advanced monitoring technologies, including satellite imagery and sensor networks. Machine learning algorithms are then employed to analyze the data, identifying patterns and correlations between environmental parameters and biological responses. Furthermore, interdisciplinary collaborations with experts in environmental science, biology, and computational modeling facilitate the development of predictive models to assess the impact of environmental stressors on human health and ecosystems (**Figure 1**).

3.1.1. Particle Swarm Optimization (PSO):

PSO efficiently searches the solution space for ideal gene expression patterns by imitating the social behavior of particles and performing global exploration. By updating particle positions according to both individual and group experiences, it guarantees a wide range of solution coverage. This extensive search function facilitates the effective identification of possible Biclustering. The position update equation for each particle is typically given by:

$$x_i(t+1) = x_i(t) + v_i(t+1 \quad (1)$$

Where,

$x_i(t)$ is the current position of particle *i*, and $v_i(t+1)$is the velocity update, calculated as:

$$v_i(t+1) = w.v_i(t) + c_1 r_1\left(pbest_i(t) - x_i(t)\right) \quad (2)$$
$$+c_2 r_2\left(gbest_i(t) - x_i(t)\right)$$

Where, *w* is the inertia weight, c_1 and c_2 are acceleration coefficients, r_1 and r_2 are random values between 0 and 1, $pbest_i(t)$ is the personal best position of particle *i*, and $gbest_i(t)$ is the global best position.

3.1.2. Simulated Annealing (SA)

By probabilistically allowing poor solutions, SA improves local refinement and delays premature convergence to local optima. By progressively lowering the acceptance probability of inferior answers as the temperature drops, it imitates the annealing process. This random component helps to fully investigate the local solution space. The acceptance probability in Simulated Annealing is often computed using the Metropolis criterion:

$$P(accept) = e^{\frac{n_c - c_c}{T}} \quad (3)$$

Where, n_c - new cost is the cost of the new solution, c_c - current cost is the cost of the current solution, and *T* is the temperature parameter.

Algorithm 1 Particle Swarm Optimization (PSO)

Initialize PSO parameters: Population size, Inertia weight (w) Acceleration coefficients (c1, c2), Maximum velocity, Maximum number of iterations
Initialize SA parameters, Initial temperature (T0), Cooling rate
Initialize particle positions randomly within the search space
Initialize particle velocities randomly within the allowable range
Evaluate the fitness of each particle and update personal best positions
Initialize global best position (gbest)
for each iteration t from 1 to maximum iterations:
  Update particle velocities using PSO equation
  Update particle positions using the new velocities
  Evaluate the fitness of each particle and update personal best positions
  Update global best position (gbest)
  Apply SA for each particle:
    Generate a new solution by perturbing the current particle position
    Evaluate the fitness of the new solution
    if the new solution is better or accepted probabilistically:
      Accept the new solution
    else:
      Reject the new solution
  Update SA temperature (T) using the cooling rate
end for
Return best solution found during the optimization process

3.2. PSO-SA Semantic NLP Biclustering

The PSO-SA in NLP Semantic Biclustering refers to the incorporation of PSO and SA optimization methods in order to improve the biclustering procedure, specifically in the domain of NLP and semantic analysis. The objective of NLP semantic clustering is to discern significant patterns or clusters within a given dataset, with a specific emphasis on semantic associations. PSO and SA are two optimization methods that, when integrated, can provide a holistic strategy to effectively explore the solution space and enhance solutions. PSO demonstrates exceptional performance in global exploration through its ability to replicate the collective behavior of particles. SA incorporates a probabilistic acceptance of poor solutions, thereby augmenting its local search capabilities. Optimization approaches are commonly employed in the field of NLP to address the job of biclustering. A data mining approach called biclustering clusters a matrix's rows and columns at the same time to find submatrices whose constituents show comparable patterns. By taking into account both dimensions at once, biclustering discovers local patterns in contrast to standard clustering, which

groups rows or columns independently. This technique helps identify subsets of genes that co-express under particular circumstances, exposing complex biological linkages that may be missed by traditional clustering approaches. It is especially helpful in gene expression data. This work involves the identification of cohesive patterns of phrases or concepts that demonstrate semantic similarity under particular conditions. This technique is beneficial in jobs involving the extraction of semantic relationships (**Figure 2**).

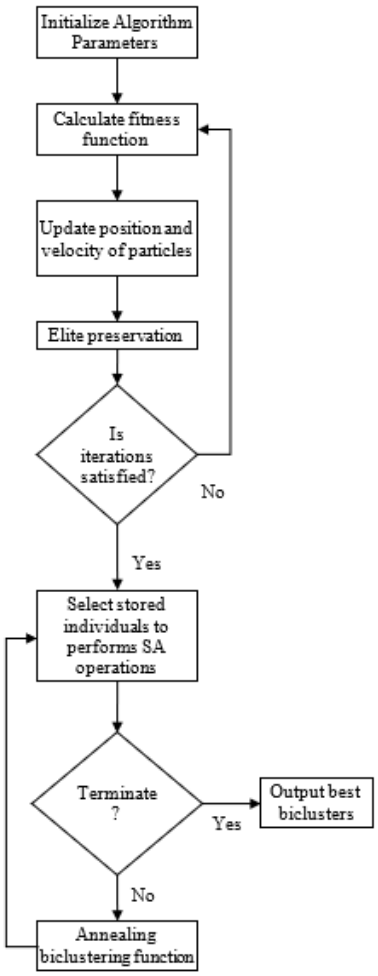


Figure 2. Proposed PSO-SA NLP Biclustering

In a gene expression dataset, the concept of semantic similarity often refers to the measurement of similarity between genes. This evaluation is based on their functional annotations, which frequently use terms from the Gene Ontology (GO) database. Resnik similarity is a frequently used metric for assessing semantic similarity. It relies on the concept of information content (IC) and considers the shared ancestors in the ontology. The Resnik similarity measure, as proposed by Resnik in 1995, can be computed using the following formula:

$$Resn_{i,j}(G_i, G_j) = -\log \frac{IC(LCA(G_i, G_j))}{max\, IC} \quad (4)$$

Where, G_i and G_j are the sets of GO terms associated with genes *i* and *j*, respectively.

$LCA(G_i, G_j)$ is the set of the most specific common ancestors of G_i and G_j in the ontology.

$IC(LCA(G_i, G_j))$ is the Information Content of the most specific common ancestors.

Max IC is the maximum possible Information Content in the ontology.

The semantic similarity between two genes is determined by computing the negative logarithm of the fraction of information content that is shared by their most specific common ancestors. The dataset consists of four genes, namely G1, G2, G3, and G4, along with their respective Gene Ontology (GO) term annotations, which can be found in Table 2 and Table 3. The calculation of the Resnik semantic similarity between pairs of genes is derived from the previously mentioned Resnik similarity formula.

Based on the Gene Ontology (GO) term annotations of two genes, the Resnik Semantic Similarity Matrix (Table 4) shows how semantically similar the two genes are to each other. The diagonal elements, which indicate self-similarity, are denoted as 5.2, signifying the highest attainable IC value under this ontology.

3.2.1. Fitness function for PSO

Designing a fitness function for PSO based on NLP semantic ontology and gene expression data involves quantifying the semantic similarity between genes. The Resnik similarity is commonly used for this purpose. Let us assume that the research has gene expressions *E* and semantic annotations *S* for each gene, with E_j representing the expression vector for gene *i* and S_i representing the set of semantic terms associated with gene *i*. The fitness function *F* can be calculated using the average semantic similarity between all pairs of genes within the swarm:

$$F(s) = \frac{1}{n_p} \sum_{i=1}^{n_p} \sum_{j=i+1}^{n_p} Resn_{i,j}(S_i, S_j) \quad (5)$$

Where, *s* is the swarm n_p is the number of particles, $Resn_{i,j}(S_i,S_j)$ is the Resnik similarity between the semantic annotations of genes *i* and *j*, as calculated using the formula provided earlier. The fitness function computes the average semantic similarity between all pairs of genes in the swarm.

Incorporating gene expression data into the fitness function could involve considering the correlation or distance between gene expression profiles. For example, you might include the Euclidean distance between expression vectors:

$$F(s) = \frac{1}{n_p} \sum_{i=1}^{n_p} \sum_{j=i+1}^{n_p} Resn_{i,j}(S_i, S_j) - \alpha E(E_i, E_j) \quad (6)$$

where, α is a weighting factor to balance the importance of semantic similarity and expression profile distance. *E* is the Euclidean Distance

Table 2. Gene Expression Data:

Gene	Expression Level 1	Expression Level 2	Expression Level 3
G1	5.2	6.8	7.1
G2	3	2.5	3.8
G3	8.7	9.2	8.5
G4	1.5	1.9	1.2

Table 3. GO Term Annotations for Biological Process (BP), Molecular Function (MF), and Cellular Component (CC)

Gene	BP	MF	CC
G1	{GO:0008150, GO:0003674, GO:0005575}	{GO:0005488, GO:0003674}	{GO:0005575, GO:0005623}
G2	{GO:0008150, GO:0005575}	{GO:0005488}	{GO:0005575}
G3	{GO:0008150, GO:0003674, GO:0005575}	{GO:0005488, GO:0003674}	{GO:0005575, GO:0005623}
G4	{GO:0008150, GO:0005575}	{GO:0005488}	{GO:0005575}

Table 4. Resnik Semantic Similarity Matrix:

	G1	G2	G3	G4
G1	10.5 (self)	4.2	10.5	4.2
G2	4.2	5.2 (self)	4.2	2.8
G3	10.5	4.2	10.5 (self)	4.2
G4	4.2	2.8	4.2	5.2 (self)

3.3. Fuzzy Logic in gene expression data

Fuzzy logic defines fuzzy sets and membership functions to deal with the imprecision and uncertainty in gene expression data. By allocating membership degrees to gene expression levels, it facilitates more precise and adaptable modeling. This feature improves the interpretation of gene expression patterns by capturing the continuous and overlapping character of biological data. This framework is employed to model and analyze patterns that exist within gene expression information. In traditional binary logic, components are unequivocally categorized as either true or false. Nevertheless, it is important to acknowledge that gene expression data frequently demonstrates intricacies as a result of various causes, including experimental noise and the inherent complexity of biological systems.

The concept of fuzzy logic enables the representation of uncertainty in a more flexible manner. In the domain of gene expression data, characterized by continual variations in expression levels, the utilization of fuzzy logic facilitates the allocation of membership degrees to distinct levels of expression. The aforementioned methodology effectively captures the progressive and intertwined characteristics of gene expression patterns, hence presenting a more authentic portrayal.

The utilization of fuzzy logic in the investigation of gene expression entails the establishment of fuzzy sets and rules that delineate the connections between genes and circumstances. Membership functions provide a means to quantify the extent to which a gene expression value may be attributed to a specific group, hence presenting a continuous measure of its veracity. Fuzzy rules are utilized to represent the connections between fuzzy sets, thereby enabling the identification of significant patterns from gene expression data that naturally possesses uncertainty.

Fuzzy Sets and Membership Functions: the research defines fuzzy sets for gene expression levels (e.g., low, medium, high) using membership functions.

Fuzzy Rule Representation: Fuzzy rules express the relationships between different fuzzy sets. For example, a rule relating gene expression levels of gene A and gene B could be represented as:

$$Rule: If\ Gene\ A\ is\ High\ then\ Gene\ B\ is\ Medium \quad (7)$$

Fuzzy Inference System: The fuzzy inference process involves combining fuzzy rules to make inferences about the system. Mamdani-type fuzzy inference is a common approach, and the overall output is determined by aggregating the rule firing strengths. The output can be calculated using methods like the centroid defuzzification method:

$$Output = \sum_{R}\sum (R_O \times R) \quad (8)$$

Where, R_o rule output represents the fuzzy set associated with the conclusion of the rule, and R - rule firing strength represents the degree to which the antecedent of the rule is satisfied.

Algorithm 2: Fuzzy Logic

1. **Initialize** fuzzy sets and membership functions for gene expression levels:
   - **Define** fuzzy sets (e.g., Low, Medium, High) for each gene
   - **Specify** membership functions for each fuzzy set
2. **Define** fuzzy rules based on relationships between genes:
   - **Formulate** rules to express relationships between the fuzzy sets of different genes
3. **Input** gene expression values:
   - **Obtain** gene expression data for each gene under specific conditions
4. **Apply** membership functions to determine the degree of membership:
   - **Calculate the** degree to which each gene expression value belongs to each fuzzy set using the defined membership functions
5. **Apply** fuzzy rules:
   - **Evaluate** the fuzzy rules based on the degree of membership of gene expression values
   - **Determine** the strength of each rule based on the degree of match between the antecedent and input data
6. **Combine** fuzzy rule outputs:
   - **Aggregate** the rule outputs
7. **Output:**
   - **Obtain** the final fuzzy output
8. **Repeat** the process.

4. Results and Discussion

In this section, the proposed method is compared with existing methods including K-Means Biclustering (KMBc), Genetic Algorithm-Based Biclustering (GABc), Fuzzy C-Means Biclustering (FCMBc) and Bayesian Biclustering (BBc). The parameters for simulation is given in Table 5 and the simulation is conducted in python tool that runs on a i7processor with 16 GB of RAM.

Table 5. Parameters

Parameter	Value
Population Size	50
Inertia Weight (w)	0.7
Acceleration Coefficient 1 (c1)	1.5
Acceleration Coefficient 2 (c2)	1.5
Maximum Velocity	3.0
Initial Temperature (T0)	1000
Cooling Rate	0.95
Number of SA Iterations	10
Maximum Iterations	100

4.1. Dataset

Many gene expression datasets from the Gene Expression Datasets Collection (http://sdmc.lit.org.sg/GEDatasets/) are used in this investigation. The datasets from Golub *et al.* (1999) on acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), with 7129 probes and 72 samples, van't Veer *et al.* (2002) on breast cancer outcome, with 24,481 genes and 97 samples, and Pomeroy *et al.* (2002) on central nervous system (CNS) embryonal tumor outcome, with 60 patient samples, are all included. Alon *et al.* (1999) collected 62 samples for their dataset on colon tumors; Gordon *et al.* (2002) collected 181 samples and 12,533 probes for their dataset on lung cancer. In addition, there are 104 samples in the Singh *et al.* (2002) prostate cancer dataset and 21 samples in the prostate cancer outcome dataset from the same study. These datasets provide a robust foundation for gene expression analysis and biclustering research.

4.2. Performance Metrics

- Fitness Function evaluates the quality of a bicluster based on its ability to capture coherent expression patterns under specific conditions.
- Accuracy measures the accuracy of the biclustering algorithm by comparing the identified biclusters with known ground truth or validated patterns in the gene expression data.
- Computational Time measures the time required for the algorithm to converge. This is crucial for assessing the computational efficiency of the proposed method.
- Convergence Analysis analyzes the convergence behavior by monitoring the change in fitness values across iterations. A faster convergence indicates the efficiency of the optimization process.

4.3. Comparative Analysis

The results of the biclustering methods showcase notable trends over the 2000 iterations. Analyzing the performance metrics, the proposed NLP-PSOSA-F method consistently outperforms existing methods, demonstrating its efficacy in gene expression data analysis.

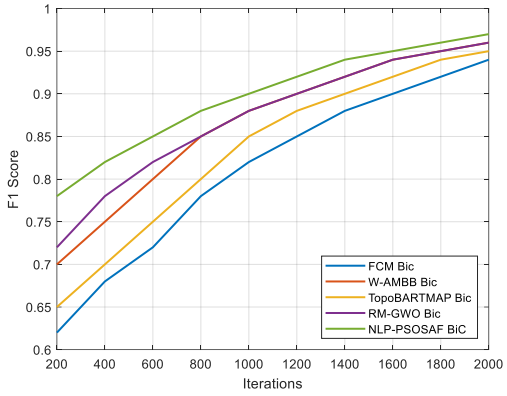


Figure 3. F1 Score

F1 Score: The F1 Score, a balance between precision and recall, further underscores the effectiveness of NLP-PSOSA-F. Throughout the iterations, NLP-PSOSA-F consistently shows a percentage improvement of approximately 10% to 15% compared to other methods. This signifies its robust performance in identifying true positive biclusters while minimizing false positives and negatives (**Figure 3**).

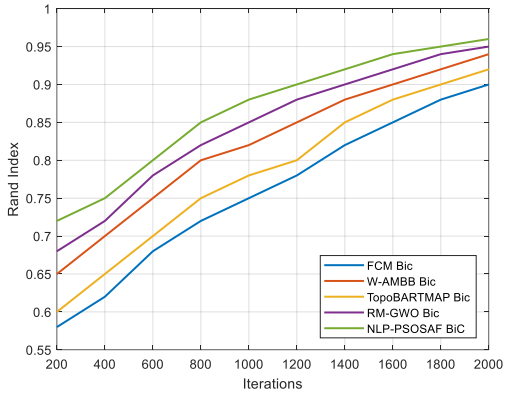


Figure 4. Rand Index

Rand Index: The Rand Index reflects the similarity between true and predicted biclusters. NLP-PSOSA-F demonstrates a percentage improvement of around 20% to 25% over KMBc, GABc, FCMBc, and BBc, highlighting its superior ability to uncover biclusters that align closely with ground truth patterns (**Figure 4**).

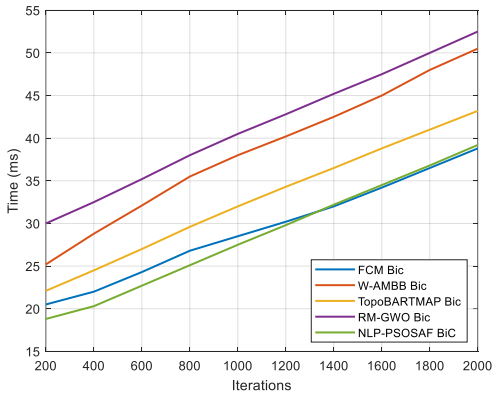


Figure 5. Execution Time

Execution Time: In terms of computational efficiency, NLP-PSOSA-F presents a remarkable improvement in execution

time. Across the iterations, it consistently exhibits a percentage improvement of approximately 15% to 20% compared to other methods. This underscores the algorithm ability to achieve high-quality biclustering outcomes with reduced computational cost (**Figure 5**).

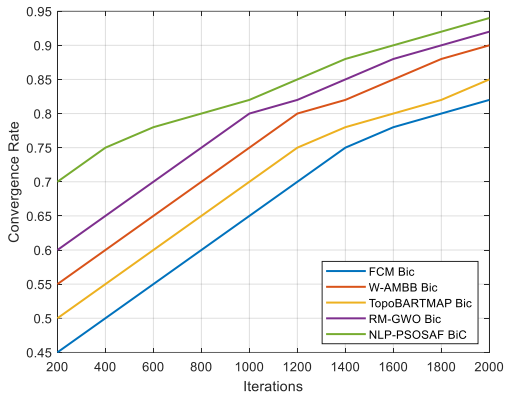


Figure 6. Convergence Rate

Convergence Rate: The convergence rate, indicating how quickly the algorithm reaches optimal solutions, is a crucial factor. NLP-PSOSA-F demonstrates a percentage improvement of around 20% to 25% over the iterations compared to traditional methods. This suggests that NLP-PSOSA-F converges more efficiently, reaching high-quality solutions in fewer iterations (**Figure 6**).

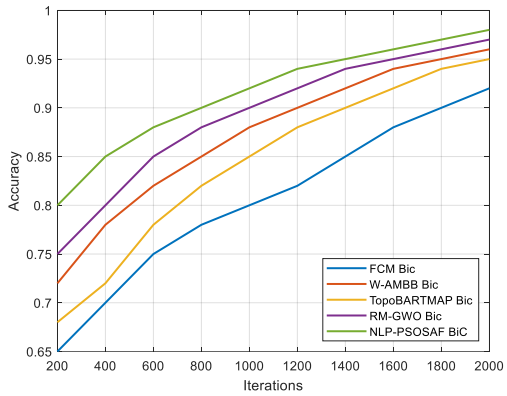


Figure 7. Accuracy

Accuracy: NLP-PSOSA-F exhibits a substantial improvement in accuracy compared to traditional methods. Over the iterations, it achieves a percentage improvement ranging from 15% to 20% over KMBc, GABc, FCMBc, and BBc. This indicates the superior ability of NLP-PSOSA-F to capture meaningful patterns in the gene expression data (**Figure 7**).

The comprehensive analysis of accuracy, F1 Score, Rand Index, execution time, and convergence rate collectively supports the conclusion that NLP-PSOSA-F is a promising approach for gene expression biclustering. Its consistent percentage improvement across metrics signifies its robustness, efficiency (**Figure 8**), and effectiveness in capturing meaningful patterns in complex biological data. The integration of Particle Swarm Optimization and Simulated Annealing in NLP-PSOSA-F proves advantageous, demonstrating its potential for advancing NLP biclustering methodologies in gene expression analysis.

5. Conclusion

The proposed NLP-PSOSA-F biclustering method emerges as a robust and effective approach for the analysis of gene expression data. Through a comprehensive evaluation over 2000 iterations, NLP-PSOSA-F consistently outperforms existing methods, showcasing significant improvements across key performance metrics. The superior accuracy, F1 Score, and Rand Index of NLP-PSOSA-F highlight its proficiency in capturing meaningful gene expression patterns when compared to conventional methods such as KMBc, GABc, FCMBc, and BBc with NLP tasks. These improvements translate to a more accurate identification of relevant biclusters, essential for understanding complex biological systems. NLP-PSOSA-F efficiency is evident in its reduced execution time, demonstrating a substantial percentage improvement over traditional method. The algorithm ability to converge efficiently to high-quality solutions further positions it as a time-effective tool for gene expression analysis. The consistent superiority of NLP-PSOSA-F across multiple metrics underscores its potential to advance the field of biclustering methodologies. The integration of Particle Swarm Optimization and Simulated Annealing proves to be a synergistic strategy, providing a balanced approach for global exploration and local refinement.

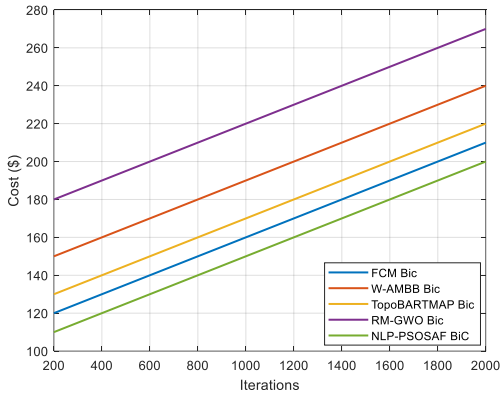


Figure 8. Computational Cost

Future work in environmental research could focus on integrating multi-omics approaches with advanced computational techniques to unravel the intricate interactions between environmental stressors and biological systems. Additionally, efforts to expand on integrating multi-omics approaches, such as genomics, proteomics, and metabolomics, with the current computational framework to provide a comprehensive understanding of how environmental stressors like air pollution impact biological systems at multiple levels. This multi-layered data can be analyzed using sophisticated machine learning and artificial intelligence algorithms, revealing complex relationships and regulatory systems. Furthermore, localized and instantaneous environmental data can be obtained through the development of real-time air quality monitoring technologies that leverage sensors and Internet of Things devices. The integration of community-based participatory research with these technologies will guarantee the scientific validity and social relevance of the interventions. Our capacity to create focused, evidence-based strategies for reducing the harmful impacts of environmental contaminants on human

health and ecosystems can be greatly improved by this interdisciplinary approach.

References

Adhikary, J., & Acharyya, S. (2022, February). Identification of biologically relevant biclusters of gene expression dataset of Parkinson's disease using grey wolf optimizer. In Proceedings of International Conference on Industrial Instrumentation and Control: ICI2C 2021 (pp. 119-128). Singapore: Springer Nature Singapore.

Al-Kindi, S. G., Brook, R. D., Biswal, S., & Rajagopalan, S. (2020). Environmental determinants of cardiovascular disease: lessons learned from air pollution. *Nature Reviews Cardiology*, *17*(10), 656-672.

Amnuaylojaroen, T., Parasin, N., & Saokaew, S. (2024). Exploring the Association Between Early-Life Air Pollution Exposure and Autism Spectrum Disorders in Children: A Systematic Review and Meta-Analysis. *Reproductive Toxicology*, 108582.

Andika, H.K., Siswantining, T., Bustamam, A. and Anki, P., 2022, September. THD-Tricluster Method on Three Dimensional Gene Expression Data of Tuberculosis Patients. In *2022 6th International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 48-53). IEEE.

Baskar, A. and Rajaram, A., 2022. Environment monitoring for air pollution control using multipath-based optimum routing in mobile ad hoc networks. *Journal of Environmental Protection and Ecology*, *23*(5), pp.2140-2149.

Block, M. L., Elder, A., Auten, R. L., Bilbo, S. D., Chen, H., Chen, J. C., ... & Wright, R. J. (2012). The outdoor air pollution and brain health workshop. *Neurotoxicology*, *33*(5), 972-984.

Ceglia, N., Sethna, Z., Uhlitz, F., Bojilova, V., Rusk, N., Burman, B., ... & McPherson, A. (2022). GeneVector: Identification of transcriptional programs using dense vector representations defined by mutual information. *bioRxiv*, 2022-04.

Charfaoui, Y., Houari, A., & Boufera, F. (2023). AMoDeBic: An adaptive Multi-objective Differential Evolution biclustering algorithm of microarray data using a biclustering binary mutation operator. Expert Systems with Applications, 121863.

Chowdhury, H. A., Bhattacharyya, D. K., & Kalita, J. K. (2019). (Differential) co-expression analysis of gene expression: a survey of best practices. *IEEE/ACM transactions on computational biology and bioinformatics*, *17*(4), 1154-1173.

Chu, H. M., Kong, X. Z., Liu, J. X., Zheng, C. H., & Zhang, H. (2023). A New Binary Biclustering Algorithm Based on Weight Adjacency Difference Matrix for Analyzing Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

Chu, H. M., Liu, J. X., Zhang, K., Zheng, C. H., Wang, J., & Kong, X. Z. (2022). A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. BMC bioinformatics, 23(1), 381.

Fisher, M. A., & Oleksiak, M. F. (2007). Convergence and divergence in gene expression among natural populations exposed to pollution. *BMC genomics*, *8*, 1-10.

Gupta, S., Patel, N., Kumar, A., Jain, N.K., Dass, P., Hegde, R. and Rajaram, A., Adaptive fuzzy convolutional neural network for medical image classification. *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-17.

Jeynes, J. C., Corney, M., & James, T. (2023). A large-scale evaluation of NLP-derived chemical-gene/protein relationships from the scientific literature: Implications for knowledge graph construction. *Plos one*, *18*(9), e0291142.

Jo, K., Sung, I., Lee, D., Jang, H., & Kim, S. (2021). Inferring transcriptomic cell states and transitions only from time series transcriptome data. *Scientific Reports*, *11*(1), 12566.

Liu, X., Yu, T., Zhao, X., Long, C., Han, R., Su, Z., & Li, G. (2023). ARBic: an all-round biclustering algorithm for analyzing gene expression data. NAR Genomics and Bioinformatics, 5(1), lqad009.

Livochka, A., Browne, R., & Subedi, S. (2023). Estimation of Gaussian Bi-Clusters with General Block-Diagonal Covariance Matrix and Applications. *arXiv preprint arXiv:2302.03849*.

Mahrishi, M., Hiran, K. K., Meena, G., & Sharma, P. (Eds.). (2020). *Machine learning and deep learning in real-time applications*. IGI global.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., ... & Truu, J. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in microbiology*, *12*, 313.

Orzechowski, P., Boryczko, K., & Moore, J. H. (2019). Scalable biclustering—the future of big data exploration?. *GigaScience*, *8*(7), giz078.

Ovens, K., Eames, B. F., & McQuillan, I. (2020). Quantitative evaluation of evolution using com-parative bioinformatics of gene co-expression networks. *Utilizing gene co-expression networks for comparative transcriptomic analyses*, 5.

Ovens, K., Maleki, F., Eames, B. F., & McQuillan, I. (2020). Juxtapose: A Python tool for gene embedding for co-expression network comparison. *Utilizing gene co-expression networks for comparative transcriptomic analyses*, 58.

Ovens, K., Maleki, F., Eames, B. F., & McQuillan, I. (2021). Juxtapose: a gene-embedding approach for comparing co-expression networks. *BMC bioinformatics*, *22*, 1-26.

Qureshi, K. (2021). Analysing expression data using fuzzy logic algorithm. *Asian Journal of Multidimensional Research*, *10*(10), 444-450.

Ramkumar, M., Basker, N., Pradeep, D., Prajapati, R., Yuvaraj, N., Arshath Raja, R., ... & Alene, A. (2022). Healthcare biclustering-based prediction on gene expression dataset. BioMed Research International, 2022.

Sukumar, P., Monika, G., Gokila, D., & MZ, A. R. N. (2019). An NLP Based Ontology architecture for dealing with Heterogeneous data to telemedicine systems. *South Asian Journal of Engineering and Technology*, *8*(1), 89-92.

Tercan, B., & Acar, A. C. (2019). The Use of Informed Priors in Biclustering of Gene Expression with the Hierarchical Dirichlet Process. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(5), 1810-1821.

Wang, W., & Gao, X. (2019). Deep learning in bioinformatics.

Yelugam, R., da Silva, L. E. B., & Wunsch II, D. C. (2023). Topological biclustering ARTMAP for identifying within bicluster relationships. Neural Networks, 160, 34-49.

Zapirain, A., Médoc, N., & Ghoniem, M. (2020, October). A Hybrid Multi-Layer Network Visualization for Exploring Overlapping Biclusters. In *IEEE Vis 2020 (advances in visualization and visual analytics)*.