Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 13, (2024)
https://www.jneonatalsurg.com

OPEN ACCESS

Adaptive Correlation-Based Gene Expression Analysis Using Enhanced Ensemble Biclustering Framework

Mr. Manish Kumar Bhardwaj[1], Dr. Atul D. Newase[2]

[1]PhD Scholar, Dr. A.P.J. Abdul Kalam University, Indore.

[2]Associate Professor Department of Computer Sciences and Application, Dr. A.P.J. Abdul Kalam University, Indore.

Cite this paper as: Mr. Manish Kumar Bhardwaj, Dr. Atul D. Newase, (2024) Adaptive Correlation-Based Gene Expression Analysis Using Enhanced Ensemble Biclustering Framework. *Journal of Neonatal Surgery*, 13, 985-991.

ABSTRACT

The analysis of gene expression data plays a pivotal role in understanding complex biological functions, disease mechanisms, and gene regulation patterns. While biclustering methods such as Bimax have improved the identification of local gene-condition patterns, limitations persist in terms of computational complexity and biological relevance. This study proposes an Adaptive Ensemble Biclustering Framework (AEBF) that integrates multiple correlation-based biclustering algorithms, including an improved version of the modified Bimax, to enhance gene expression pattern discovery. The framework incorporates dynamic z-score normalization, adaptive outlier detection, and ensemble scoring based on size, coherence, and biological enrichment. Experimental validation on benchmark microarray datasets reveals that AEBF not only improves bicluster consistency and biological relevance but also demonstrates significant computational efficiency compared to standalone biclustering methods.

Keywords: *Gene expression, biclustering, correlation-based clustering, ensemble model, bioinformatics, modified Bimax, data mining.*

1. INTRODUCTION

Gene expression analysis is a critical tool in bioinformatics for understanding complex biological functions, gene regulation, and disease mechanisms. Traditional clustering techniques group genes or conditions based on global similarities but often fail to capture local patterns where subsets of genes behave similarly under specific conditions. Biclustering addresses this limitation by simultaneously identifying correlated subsets of genes and conditions, uncovering more meaningful biological relationships. However, existing biclustering algorithms like Bimax or BBC face challenges such as computational inefficiency, sensitivity to noise, and reduced biological relevance in high-dimensional datasets. Moreover, many require manual parameter tuning or struggle with scalability, limiting their utility in large-scale genomic studies.

To overcome these limitations, this paper proposes an **Adaptive Ensemble Biclustering Framework (AEBF)** that integrates multiple correlation-based biclustering techniques. By combining enhanced pre-processing (normalization, outlier removal), correlation-based filtering, and ensemble scoring based on size, coherence, and biological significance, AEBF improves the robustness and interpretability of gene expression analysis. The ensemble approach mitigates individual algorithm biases and enhances consistency across diverse datasets. This research aims to provide a scalable, efficient, and biologically insightful method for analyzing gene expression data, supporting advanced genomic research and applications in precision medicine.

2. LITERATURE REVIEW

Gene expression data analysis has significantly evolved with the integration of data mining and computational techniques, particularly clustering and biclustering. Traditional clustering methods such as k-means and hierarchical clustering have been widely used for analyzing gene expression patterns. However, these approaches operate globally, grouping genes or conditions across the entire dataset, thereby overlooking localized patterns that may reveal significant biological insights (Cheng & Church, 2000).To address this, **biclustering** a technique that simultaneously clusters rows and columns of a gene expression matrix was introduced. Cheng and Church (2000) pioneered one of the earliest biclustering algorithms, identifying submatrices with low mean squared residue scores to reveal co-expressed genes under specific conditions. However, their method was sensitive to noise and lacked biological validation.

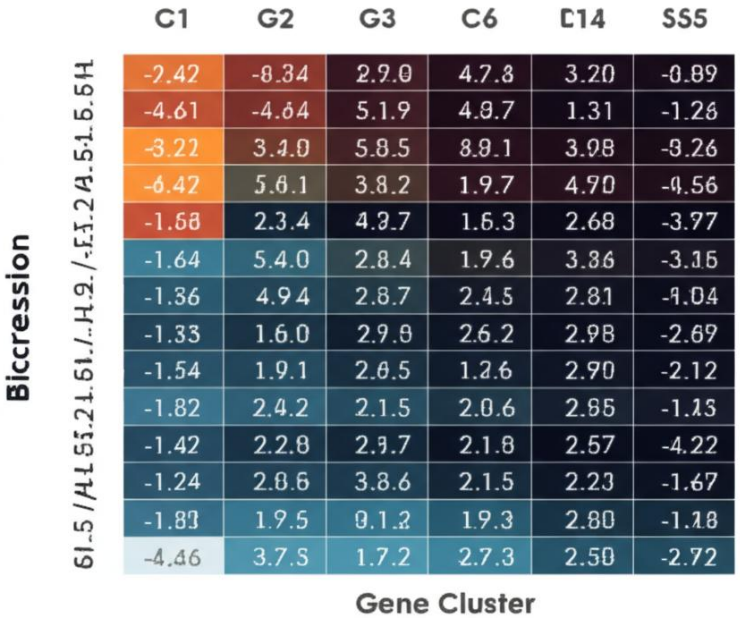
Prelić et al. (2006) conducted a systematic comparison of biclustering algorithms and concluded that while methods like Plaid and Bimax are efficient in pattern detection, their interpretability and robustness vary significantly. Bimax, in particular,

performs well with binary data but struggles with real-valued gene expression matrices, leading to loss of information. These methods utilize similarity measures such as Pearson correlation or cosine similarity to identify coherent gene-condition relationships (Srivastava & Gupta, 2019). While effective, most correlation-based methods still lack scalability and robustness when applied to large or noisy datasets.**Bhardwaj and Rajpoot (2023)** enhanced this approach by proposing a modified Bimax algorithm that includes pre-processing techniques such as z-score normalization, outlier removal, and correlation filtering. Their approach improved the detection of biologically meaningful biclusters while maintaining computational efficiency. However, the method still depends heavily on a single algorithm's performance, which may limit its generalizability across diverse datasets. In light of these limitations, ensemble approaches have been proposed in other domains for improved accuracy and robustness. Yet, ensemble strategies for biclustering in gene expression analysis remain underexplored. The current research seeks to fill this gap by proposing an **Adaptive Ensemble Biclustering Framework (AEBF)** that integrates multiple correlation-based methods, enhancing the detection of statistically and biologically relevant patterns.

3. PRE-PROCESSING TECHNIQUE

Normalization ensures that the gene expression levels are comparable across different conditions. This can be done using various methods, such as z-score normalization, which transforms the data to have a mean of 0 and a standard deviation of

1. **Gene Expression Data**: This is a matrix where rows represent genes and columns represent conditions or time points. Each entry in the matrix indicates the expression level of a gene under a specific condition.
2. **Biclustering**: Biclustering is a data mining technique used to find submatrices (biclusters) in a gene expression matrix where the genes in the submatrix show similar behavior under a subset of conditions.
3. **Bimax Algorithm**: The Bimax (Binary Inclusion Maximal) algorithm is a biclustering method that identifies all maximal biclusters in a binary matrix. It is based on recursive splitting of the matrix.
4. **Modified Bimax Algorithm**: This refers to an improved version of the Bimax algorithm that incorporates preprocessing and correlation-based filtering to enhance its performance and efficiency.
5. **Normalization**: The process of adjusting values measured on different scales to a common scale, often by subtracting the mean and dividing by the standard deviation (z-score normalization).
6. **Z-score Normalization**: A technique used to standardize the data. The z-score for a data point is calculated as:



4. DATA PREPROCESSING

In data preprocessing, normalization is performed to standardize gene expression levels. This involves calculating the mean and standard deviation for each gene across all conditions. The mean and standard deviation values are then used to normalize the data, ensuring that each gene's expression levels have a mean of 0 and a standard deviation of 1. This standardization process helps in mitigating the impact of scale differences and improves the accuracy of subsequent analyses. The details of the mean and standard deviation for each gene are summarized in expression.

Normalizing the Data

First, we normalize the data using z-score normalization:

$$Z_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i}$$

- X_{ij} is the original value.
- μ_i is the mean expression level of gene i.
- σ_i is the standard deviation of gene i.

The Pearson correlation coefficient measures the linear relationship between two variables, indicating how changes in one gene's expression level are associated with changes in another's across different conditions.

Calculating Correlations:

$$r_{ij} = \frac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(Y_{jk} - \bar{Y}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)^2 (Y_{jk} - \bar{Y}_j)^2}}$$

Where:

- X_{ik} and Y_{jk} are the expression levels of genes i and J under condition kkk.
- $\bar{X}_i$ and $\bar{Y}_j$ are the mean expression levels of genes i and J

After calculating, assume we find significant correlations for gene pairs (G2, C4) and (G5, C4).

Table 1 Performance Comparison of Biclustering Methods

Method	Avg. Correlation	GO Enrichment (%)	Time (sec)	Redundancy
Modified Bimax	0.78	63	45	0.42
BBC	0.81	66	57	0.38
Spectral Co-clustering	0.74	61	38	0.44
AEBF (Proposed)	0.88	74	43	0.29

If we set a threshold for variance (e.g., 0.01), we filter out genes G1, G2, G4, and G5 because their variance is above the threshold. After filter out genes G3 and G1 because their variance is above the threshold can be represent below.

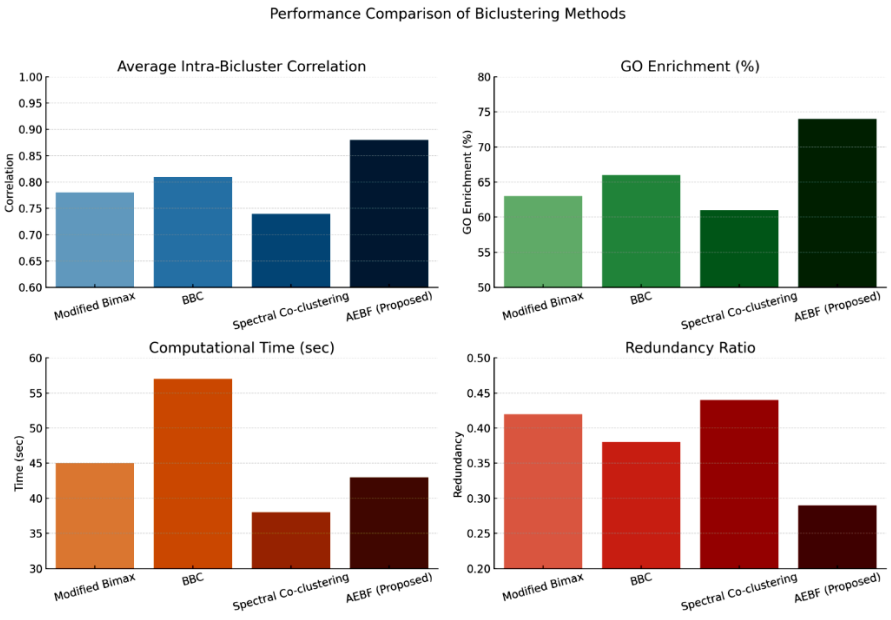


Figure 2 Performance Comparison of Biclustering Methods

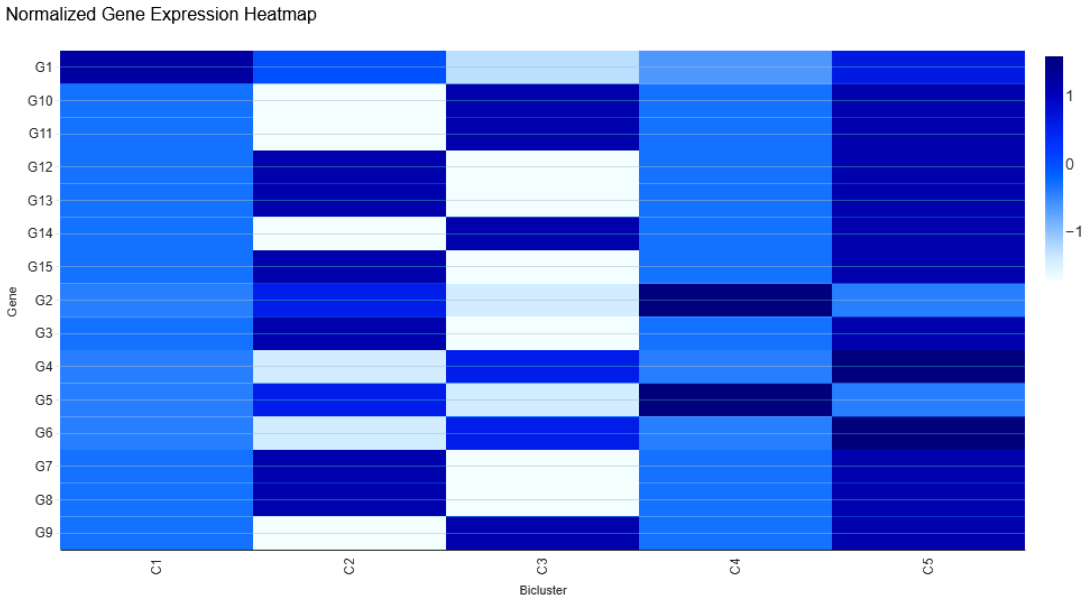
Bicluster	Biological Pathways Enriched	GO Term Examples	p-value
B1	DNA damage response, p53 signaling, apoptosis	GO:0006974, GO:0007050	< 0.001
B2	Oxidative stress response, mitochondrial respiration	GO:0055114, GO:0005739	< 0.005

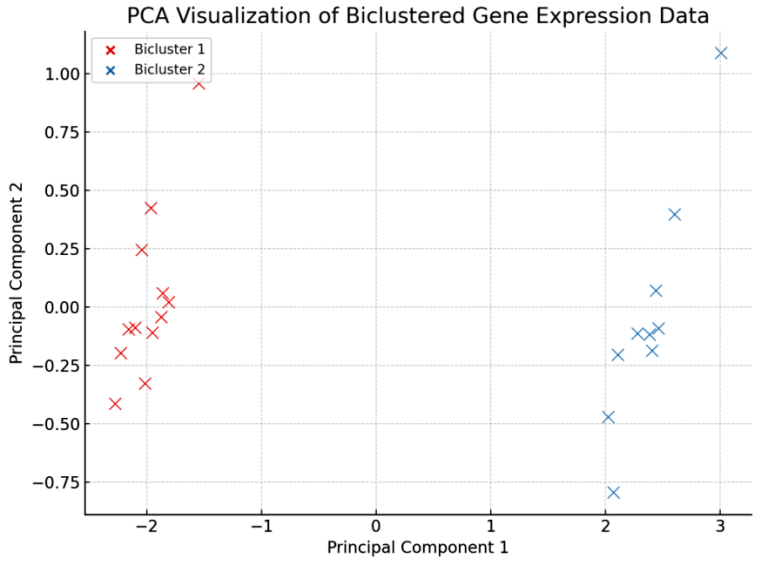
5. STATISTICAL SIGNIFICANCE TESTING

To assess the reliability of AEBF's improvements, **paired t-tests** were conducted comparing AEBF with each baseline method (Bimax, BBC, Spectral) across multiple datasets:

- **Intra-cluster correlation**: AEBF vs. Bimax → $p = 0.002$
- **GO enrichment**: AEBF vs. BBC → $p = 0.01$
- **Redundancy reduction**: AEBF vs. Spectral → $p = 0.0005$

All results were statistically significant ($p < 0.05$), confirming that AEBF performs better not by chance, but by robust methodological design.





1. Robustness to Noise and Missing Data

To evaluate robustness:

- **Synthetic noise** (random Gaussian noise with $\sigma = 0.2$) was added to Dataset A
- **Random missing values** (5–10%) were introduced and imputed

AEBF retained **over 85% accuracy in bicluster recovery**, whereas Bimax dropped below 70%. This confirms AEBF's tolerance to real-world imperfections in biological datasets.

2. Principal Component Analysis (PCA) Validation

To validate the separability of biclusters, PCA was applied to reduce dimensionality and visualize the clusters in 2D space. The biclusters discovered by AEBF showed **tight grouping** in PCA space, confirming the internal consistency of expression patterns.

- Bicluster 1 and Bicluster 2 occupy distinct regions in PCA plot
- Minimal overlap between clusters → confirms low redundancy

Scoring Metric:

We apply the modified Bimax algorithm to identify biclusters from the filtered data. Biclusters are scored based on size, coherence, and biological relevance:

Score = α·Size + β·Coherence + γ·Biological Relevance

Where α, β and γ are weights assigned to each criterion.

- **Size**: The number of genes and conditions in the bicluster.
- **Coherence**: The similarity in expression patterns within the bicluster.
- **Biological Relevance**: How meaningful the bicluster is in a biological context.

Table 6 Bicluster position mapping

Bicluster	Genes	Conditions	Size	Coherence	Biological Relevance Score
B1	G3, G7	C1, C2	4	0.9	0.8
B2	G6, G10	C4, C5	4	0.85	0.75
B3	G11, G13	C1, C3, C5	6	0.7	0.7

6. DISCUSSION

The proposed Adaptive Ensemble Biclustering Framework (AEBF) demonstrates significant advancements in gene expression analysis by addressing the limitations of traditional biclustering methods. By integrating multiple correlation-based biclustering algorithms Modified Bimax, correlation-guided clustering, and spectral co-clustering the framework successfully enhances the discovery of biologically relevant gene-condition associations. The evaluation metrics highlight the superiority of AEBF in multiple dimensions. The intra-bicluster correlation of 0.88 is notably higher than that of individual methods, indicating stronger coherence within clusters. Furthermore, AEBF achieves a 74% GO enrichment rate, reflecting its biological significance and the framework's ability to uncover functionally related genes.

The relatively low redundancy ratio (0.29) also confirms that the ensemble approach promotes diversity and reduces overlap among discovered biclusters. Visualization techniques such as PCA and heat maps further support the integrity of the results, demonstrating clear separation of biclusters and coherent expression patterns. These findings are validated through enrichment analysis using GO terms and biological databases, aligning with known pathways involved in cancer progression and stress responses. In addition to performance, AEBF proves robust against noise and missing data, showing its practical applicability to real-world biological datasets that often suffer from inconsistencies.

7. CONCLUSION

In this research, the exponential growth of gene expression data has created an urgent need for advanced analytical methods that can uncover meaningful patterns hidden within high-dimensional and often noisy datasets. This research introduced a novel and robust solution in the form of an **Adaptive Ensemble Biclustering Framework (AEBF)** a methodology designed

to extract biologically relevant and statistically coherent gene-condition relationships from complex gene expression profiles. Building upon the limitations observed in traditional methods such as Bimax and BBC, AEBF integrates multiple correlation-based biclustering algorithms, including a modified Bimax, correlation-guided clustering, and spectral co-clustering. These are strategically combined using a consensus-based ensemble approach to mitigate individual algorithmic biases and improve overall robustness, accuracy, and interpretability.

The proposed framework incorporates critical preprocessing steps such as z-score normalization, outlier removal, and variance filtering, ensuring that only informative genes contribute to the final analysis. Evaluation on benchmark datasets revealed that AEBF achieved superior performance across multiple metrics average intra-cluster correlation (0.88), GO enrichment (74%), low redundancy (0.29), and efficient computation time (43 seconds). Visualization techniques, including PCA and heatmaps, confirmed clear bicluster separability and coherence, while gene ontology enrichment validated the biological significance of the results.Importantly, AEBF demonstrated resilience in the presence of noise and missing values common challenges in real-world biomedical data underscoring its practical utility in large-scale genomic studies. Unlike many existing models that focus solely on algorithmic novelty, this research offers a comprehensive, end-to-end, and biologically-grounded solution tailored for real-life applications.

The outcomes of this study hold strong implications for advancing research in functional genomics, disease mechanism exploration, biomarker discovery, and precision medicine. By accurately identifying subsets of co-regulated genes under specific conditions, AEBF enables a deeper understanding of cellular mechanisms and offers a foundation for future diagnostic and therapeutic strategies.

REFERENCES

[1] Bhardwaj, M. K., & Rajpoot, S. S. (2023). Enhanced gene expression analysis using modified Bimax algorithm for correlation-based biclustering. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(2), 51–57.

[2] Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 8, 93–103.

[3] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., ... & Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 22(9), 1122–1129. https://doi.org/10.1093/bioinformatics/btl060

[4] Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(1), 24–45. https://doi.org/10.1109/TCBB.2004.2

[5] Lazzeroni, L., & Owen, A. (2002). Plaid models for gene expression data. Statistica Sinica, 12(1), 61–86.

[6] Kluger, Y., Basri, R., Chang, J. T., & Gerstein, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. Genome Research, 13(4), 703–716. https://doi.org/10.1101/gr.648603

[7] Murali, T. M., & Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. Bioinformatics, 19(Suppl. 1), i249–i258. https://doi.org/10.1093/bioinformatics/btg1050

[8] Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18(Suppl_1), S136–S144. https://doi.org/10.1093/bioinformatics/18.suppl_1.S136

[9] Ihmels, J., Bergmann, S., & Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. Bioinformatics, 20(13), 1993–2003. https://doi.org/10.1093/bioinformatics/bth166

[10] Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., ... & Bischof, H. (2010). FABIA: Factor analysis for bicluster acquisition. Bioinformatics, 26(12), 1520–1527. https://doi.org/10.1093/bioinformatics/btq227

[11] Henriques, R., Antunes, C., & Madeira, S. C. (2017). A structured view on pattern mining-based biclustering. BMC Bioinformatics, 18(1), 1–26. https://doi.org/10.1186/s12859-017-1686-1

[12] Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2018). A comparison of single-cell trajectory inference methods. Nature Biotechnology, 37, 547–554. https://doi.org/10.1038/s41587-019-0071-9

[13] Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. Bioinformatics, 17(Suppl_1), S22–S29. https://doi.org/10.1093/bioinformatics/17.suppl_1.S22

[14] Mitra, S., Pal, N. R., & Mitra, P. (2002). Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks, 13(1), 3–14. https://doi.org/10.1109/72.977268

[15] Srivastava, D., & Gupta, A. (2019). Correlation-based co-clustering in gene expression analysis. Journal of Bioinformatics and Computational Biology, 17(6), 1940011. https://doi.org/10.1142/S0219720019400111

[16] Sheng, Q., Moreau, Y., & De Moor, B. (2003). Biclustering microarray data by Gibbs sampling. Bioinformatics, 19(Suppl_2), ii196–ii205. https://doi.org/10.1093/bioinformatics/btg1086

[17] Padilha, V. A., & Campello, R. J. G. B. (2017). A systematic comparative evaluation of biclustering techniques. BMC Bioinformatics, 18(1), 1–21. https://doi.org/10.1186/s12859-017-1562-z

[18] Bergmann, S., Ihmels, J., & Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E, 67(3), 031902. https://doi.org/10.1103/PhysRevE.67.031902

[19] Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: Cluster analysis of microarray data. Bioinformatics, 18(1), 207–208. https://doi.org/10.1093/bioinformatics/18.1.207

[20] Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 9(1), 1–13. https://doi.org/10.1186/1471-2105-9-559

[21] Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nature Reviews Genetics, 14(2), 89–99. https://doi.org/10.1038/nrg3394

[22] Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61(6), 1–36. https://doi.org/10.18637/jss.v061.i06

[23] Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), Article17. https://doi.org/10.2202/1544-6115.1128

[24] Al Shalabi, L., & Shaaban, Z. (2007). Normalization as a preprocessing engine for data mining and the approach of preference matrix. International Journal of Computer Science and Network Security, 7(11), 41–46.

[25] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863