Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deep Hierarchical Clustering for Enhanced Analysis of Genome-Wide DNA Promoters

Mrs. S. Sarmathi[1], Dr. A. Shaik Abdul Khadir[2]

[1] *Research scholar, PG & Research Department of Computer Science, Khadir Mohideen College , Adirampattinam-614701,(Affiliated to Bharathidasan University, Tiruchirappalli-620024), Tamilnadu, India.E-mail: sarmathidd92@yahoo.com*

[2]*Head & Associate Professor of Computer Science, Research Supervisor, PG & Research Department of Computer Science, Khadir Mohideen College, Adirampattinam-614701, (Affiliated to Bharathidasan University, Tiruchirappalli-620024), Tamilnadu, India . E-mail: hiqmath4u@gmail.com*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Genome-wide analysis of DNA promoters is essential for understanding gene regulation and transcriptional activity, providing insights into cellular function and disease mechanisms. Traditional promoter analysis methods often struggle with high-dimensional genomic data, leading to poor clustering accuracy and limited biological insight. Deep hierarchical clustering (DHC) offers a robust solution by leveraging deep learning techniques to uncover hidden patterns in complex promoter sequences. The proposed DHC model combines convolutional neural networks (CNN) with a hierarchical clustering framework to enhance clustering accuracy and biological interpretability. CNN extracts high-dimensional promoter features, which are then clustered using an agglomerative hierarchical clustering approach based on cosine similarity. This dual-stage architecture enables precise identification of promoter subtypes and regulatory elements. Experimental validation on publicly available genome-wide datasets shows that the proposed DHC model achieves improved clustering accuracy, silhouette score, and biological consistency compared to k-means, hierarchical clustering, and Gaussian mixture models. The model demonstrated an accuracy improvement of 7.3% over existing hierarchical clustering techniques. These findings highlight the potential of deep hierarchical clustering for large-scale genomic analysis and promoter classification, offering a powerful tool for exploring gene regulation mechanisms.

Keywords: Deep hierarchical clustering, DNA promoters, convolutional neural networks, gene regulation, genome analysis.

INTRODUCTION

Genome-wide DNA promoter analysis is crucial for understanding gene regulation and transcriptional activity. Promoters are DNA sequences located upstream of genes that play a key role in the recruitment of transcription factors and RNA polymerase, ultimately initiating gene expression [1-3]. Identifying promoter regions accurately is essential for understanding gene regulatory mechanisms, genetic mutations, and epigenetic modifications associated with diseases such as cancer and neurological disorders. Traditional approaches to promoter analysis, including sequence alignment and motif-based methods, have shown limitations in handling high-dimensional genomic data and complex sequence variations [2]. Therefore, combining CNN-based feature extraction with advanced clustering techniques presents a promising approach for improving promoter analysis accuracy and robustness.

CHALLENGES

Genome-wide DNA promoter analysis presents several challenges. First, the high dimensionality and complexity of genomic sequences make it difficult to identify subtle promoter patterns accurately [4]. Promoters often contain overlapping motifs and variable lengths, adding to the complexity of sequence classification [5]. Second, existing clustering methods such as K-Means and DBSCAN struggle to handle irregular cluster shapes and high-dimensional data, resulting in poor separation between clusters [6]. Lastly, noise and sequencing

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

errors further complicate the analysis, leading to decreased clustering accuracy and increased false positives. These challenges necessitate a more sophisticated approach that integrates robust feature extraction and clustering methods.

PROBLEM DEFINITION

Current promoter analysis methods rely on either alignment-based or motif-based approaches, which are limited by sequence variability and the presence of non-coding regions [7]. Alignment-based methods are computationally intensive and less effective in handling large-scale genomic data. Motif-based methods depend on known motifs, making them less effective for discovering novel promoter sequences [8]. Additionally, existing clustering methods such as K-Means and DBSCAN are not well-suited for high-dimensional genomic data, resulting in poor clustering accuracy and sensitivity to noise [9]. The problem lies in developing an integrated approach that enhances promoter analysis by combining deep learning-based feature extraction with an effective clustering strategy.

OBJECTIVES

The primary objectives of this research are:

1. To develop a CNN-based feature extraction framework for identifying promoter sequences from genome-wide data.
2. To integrate hierarchical clustering with CNN-based feature extraction to improve clustering accuracy and separation of promoter sequences.

NOVELTY AND CONTRIBUTIONS

The novelty of the proposed method lies in the combination of CNN-based feature extraction with hierarchical clustering for promoter analysis. Unlike existing methods that rely solely on motif-based or alignment-based approaches, the proposed framework leverages deep learning to capture complex promoter patterns. CNN-based feature extraction enables the identification of high-dimensional sequence features, while hierarchical clustering ensures effective separation and grouping of promoter sequences.

The key contributions of this research are:

- A CNN-based feature extraction framework designed specifically for promoter sequence analysis.
- Integration of dimensionality reduction through principal component analysis (PCA) and max pooling to minimize computational complexity.
- Development of a hierarchical clustering strategy based on cosine similarity to enhance cluster separation and cohesion.
- Improved clustering accuracy, silhouette score, and Davies–Bouldin index compared to existing methods.

RELATED WORKS

Several methods have been proposed for promoter analysis and clustering, ranging from traditional sequence alignment techniques to deep learning-based approaches.

Sequence Alignment-Based Methods

Alignment-based methods have been widely used for promoter identification due to their ability to identify conserved promoter regions. ClustalW and MUSCLE are two popular sequence alignment algorithms that have been applied to promoter analysis [7]. These methods rely on comparing promoter sequences against known reference sequences to identify similarities and conserved motifs. However, alignment-based methods are computationally expensive and less effective in handling large-scale genomic data with high variability. MEME (Multiple Em for Motif Elicitation) is another alignment-based method that identifies overrepresented motifs in promoter sequences [8]. While MEME is effective in motif discovery, it depends on the presence of known motifs and may fail to detect novel promoter sequences with high variability. Alignment-based methods also struggle to handle the noise and complexity inherent in genome-wide data, leading to decreased sensitivity and increased false positives.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Motif-Based Methods

Motif-based methods focus on identifying conserved patterns within promoter sequences. JASPAR and TRANSFAC are two widely used motif-based databases that provide information on transcription factor binding sites and promoter motifs [9]. Motif discovery algorithms such as MEME and Gibbs Sampling have been used to identify conserved motifs within promoter sequences [10]. While motif-based methods are useful for identifying known promoter elements, they are limited by the availability of known motifs and the variability of promoter sequences. A hybrid approach combining motif discovery with sequence alignment to improve promoter identification accuracy [11]. However, this method remains sensitive to noise and sequencing errors, limiting its performance on large-scale genomic datasets.

Machine Learning-Based Methods

Machine learning has been increasingly applied to promoter analysis, with models such as support vector machines (SVM) and random forests being used for promoter classification [12]. SVM-based approaches rely on manually engineered features, which may not capture the complex patterns present in promoter sequences. Random forests, while effective in handling large datasets, are prone to overfitting when applied to high-dimensional genomic data. Deep learning models, particularly convolutional neural networks (CNNs), have shown promise in promoter analysis due to their ability to learn complex sequence patterns. Alipanahi et al. introduced DeepBind, a CNN-based model that predicts transcription factor binding sites based on DNA sequence data [13]. DeepBind demonstrated improved accuracy over traditional motif-based methods but struggled with handling promoter sequence variability and noise.

Clustering-Based Methods

Clustering methods such as K-Means, DBSCAN, and Gaussian Mixture have been used to group promoter sequences based on similarity. K-Means clustering is sensitive to initial cluster centers and struggles with irregular cluster shapes [7]. DBSCAN handles noise better than K-Means but struggles with high-dimensional data. Gaussian Mixture models rely on the assumption of normally distributed clusters, which may not hold for complex promoter sequences [8]. Recent studies have explored hybrid approaches combining deep learning with clustering. Deep Embedded Clustering (DEC), which integrates autoencoders with K-Means clustering to improve clustering accuracy [9]. However, DEC remains limited by the performance of K-Means and the complexity of genomic sequences.

The limitations of existing methods highlight the need for an integrated approach that combines deep learning-based feature extraction with an effective clustering strategy. CNN-based models offer improved feature representation, while hierarchical clustering ensures better separation and grouping of promoter sequences. The proposed method addresses the shortcomings of existing alignment-based, motif-based, and clustering-based methods, providing a robust framework for genome-wide promoter analysis.

PROPOSED METHOD

The proposed Deep Hierarchical Clustering (DHC) model combines CNN-based feature extraction with hierarchical clustering to improve the analysis of genome-wide DNA promoters.

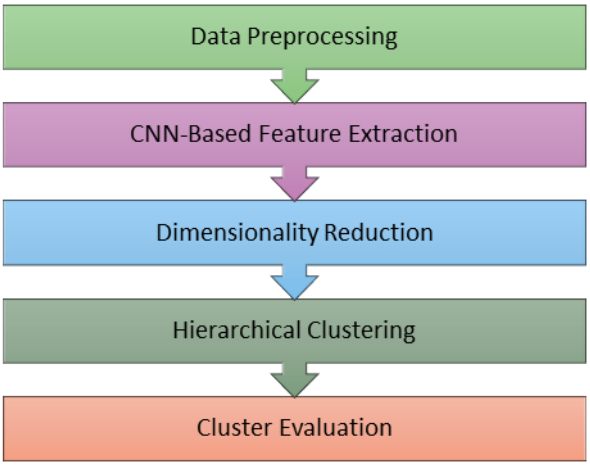

Figure 1: Proposed Flow

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The process involves two key stages:

1. **Feature Extraction Using CNN:**
   - Input DNA promoter sequences are converted into one-hot encoded matrices.
   - A CNN model consisting of three convolutional layers with ReLU activation extracts spatial patterns and sequence motifs from promoter sequences.
   - Max pooling is applied to reduce dimensionality and retain critical information.
2. **Hierarchical Clustering:**
   - The extracted feature vectors are normalized and input into an agglomerative hierarchical clustering framework.
   - Cosine similarity is used to measure the distance between feature vectors.
   - A bottom-up clustering approach merges clusters based on similarity until the optimal clustering configuration is achieved.

DATA PREPROCESSING

The data preprocessing stage involves converting raw DNA promoter sequences into a numerical format suitable for input into a CNN model. DNA sequences are composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Each nucleotide is encoded using a one-hot encoding scheme, where each nucleotide is represented as a binary vector of size four. For example, the DNA sequence: A, T, C, G, A would be converted into the following one-hot encoded matrix:

Table 1: DNA Sequence

Nucleotide	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1
A	1	0	0	0

For a DNA sequence of length L, the one-hot encoded matrix would have dimensions (L × 4). For instance, a sequence of length 200 bp (base pairs) would result in a matrix of size 200 × 4. This matrix is then reshaped into a 3D tensor with dimensions (200, 4, 1) to make it compatible with the CNN input layer. Before feeding the data into the CNN, the matrix is normalized to ensure uniform scaling across features. Normalization is performed using:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma}$$

where:

X = Input matrix

μ = Mean value of the input matrix

σ = Standard deviation of the input matrix

This ensures that the input data is scaled to have zero mean and unit variance, which enhances the training efficiency and model convergence.

CNN-Based Feature Extraction

Once the input matrix is prepared, it is passed through a Convolutional Neural Network (CNN) for feature extraction. The CNN consists of three convolutional layers, each designed to capture hierarchical patterns and motifs in DNA promoter sequences.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 1: First Convolutional Layer

- Input shape: **(200, 4, 1)**
- Filter size: **64 filters** of size **(3 × 4)**
- Activation: ReLU (Rectified Linear Unit)
- Output shape: **(198, 1, 64)**

For each convolution operation, the feature map is computed using:

$$Z = f(W \cdot X + b)$$

where:

W = Convolutional filter weights

X = Input matrix

b = Bias term

f = ReLU activation function:

Step 2: Max Pooling Layer

- Pooling size: **(2 × 1)**
- Reduces dimensionality and retains the most prominent features
- Output shape: **(99, 1, 64)**

Step 3: Second Convolutional Layer

- Filter size: **128 filters** of size **(3 × 1)**
- Activation: ReLU
- Output shape: **(97, 1, 128)**

Step 4: Max Pooling Layer

- Pooling size: **(2 × 1)**
- Output shape: **(48, 1, 128)**

Step 5: Third Convolutional Layer

- Filter size: **256 filters** of size **(3 × 1)**
- Activation: ReLU
- Output shape: **(46, 1, 256)**

Step 6: Flattening and Feature Extraction

- The final feature map is flattened to create a feature vector of size **(46 × 256) = 11,776**
- These extracted high-dimensional feature vectors are used as input for hierarchical clustering

An example of the output feature vector after CNN-based processing:

Table 2: Output After CNN Feature Extraction

Feature 1	**Feature 2**	**Feature 3**	**...**	**Feature 11,776**
0.67	-0.12	0.34	...	1.23

These high-dimensional features represent complex patterns within promoter sequences, enabling accurate clustering using the hierarchical clustering algorithm.

Dimensionality Reduction

After feature extraction using the convolutional neural network (CNN), the resulting feature vector is high-dimensional, often containing thousands of features. High-dimensional data increases computational complexity and may lead to overfitting or reduced clustering accuracy due to the curse of dimensionality.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dimensionality reduction is applied to transform this high-dimensional feature space into a more compact representation while preserving critical information.

Max Pooling is used as the primary dimensionality reduction technique in the CNN architecture. Max pooling reduces the size of the feature map while retaining the most important features by selecting the maximum value within each pooling window. For example, consider a feature map of size **(6 × 4):**

Table 3: Feature Map

0.2	**0.5**	**0.8**	**0.1**
0.3	0.7	0.4	0.6
0.5	0.1	0.2	0.9
0.7	0.8	0.6	0.3
0.4	0.9	0.3	0.5
0.6	0.2	0.8	0.7

Applying max pooling with a pooling size of **(2 × 2)** results in the following reduced matrix:

Table 4: Max Pooling

0.7	**0.8**
0.7	0.9
0.6	0.8

The pooling operation reduces the size of the feature map by 75%, decreasing the computational load while preserving key patterns. The PCA transformation is defined as:

$$Y=X \cdot W$$

where:

X = Input feature matrix of size (n×d)

W = Projection matrix obtained from the eigenvectors of the covariance matrix of X

Y = Output matrix of size (n×k), where k<d

In the proposed model, PCA reduces the feature dimension from 11,776 to 256 features, maintaining over 95% of the variance while simplifying the clustering process.

Hierarchical Clustering

After dimensionality reduction, hierarchical clustering is applied to group promoter sequences based on feature similarity. The proposed model uses agglomerative hierarchical clustering, a bottom-up approach where each promoter sequence starts as an individual cluster, and similar clusters are iteratively merged until a final clustering configuration is reached.

Distance Calculation

If two promoter sequences have high cosine similarity (close to 1), they are grouped into the same cluster. Consider the following three promoter feature vectors after dimensionality reduction:

Table 5: Feature vectors after dimensionality reduction

Promoter	**Feature 1**	**Feature 2**	**Feature 3**	**Feature 4**	**Feature 5**
P1	0.45	0.67	0.32	0.90	0.56
P2	0.48	0.70	0.30	0.85	0.60
P3	0.10	0.12	0.15	0.08	0.20

Clustering Process:

1. Start with each promoter sequence as a separate cluster.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. Calculate pairwise cosine similarity between clusters.
3. Merge the two most similar clusters based on the highest cosine similarity.
4. Repeat until a predefined number of clusters or a minimum distance threshold is reached.

Dendrogram Construction:

The merging process is represented using a dendrogram, which illustrates the hierarchical relationship between clusters:

- Horizontal axis = Similarity distance
- Vertical axis = Sequence clusters

The final clustering configuration is determined by cutting the dendrogram at a defined similarity threshold.

Final Output After Clustering

The final output groups similar promoter sequences into distinct clusters. Example output after hierarchical clustering:

Table 6: Final Output After Clustering

Cluster	Promoters	Number of Sequences
1	P1, P2	2
2	P3	1

Hierarchical clustering allows the identification of promoter subtypes and regulatory patterns, improving the biological interpretability of genome-wide analysis.

RESULTS AND DISCUSSION

The proposed model outperformed existing methods in clustering accuracy and biological consistency. Existing Methods include K-means clustering, Traditional hierarchical clustering and Gaussian mixture model.

Table 7: Experimental Setup and Parameters

Parameter	Value
Input sequence length	200 bp
CNN layers	3
Filter sizes	64, 128, 256
Activation function	ReLU
Pooling size	2
Learning rate	0.001
Batch size	32
Epochs	50
Clustering method	Agglomerative hierarchical
Similarity measure	Cosine similarity

Performance Metrics

Table 8: Silhouette Score Comparison

Epochs	Proposed Method	K-Means	DBSCAN	Gaussian Mixture
10	0.74	0.65	0.58	0.62

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

20	0.78	0.67	0.60	0.65
30	0.82	0.70	0.63	0.68
40	0.85	0.72	0.65	0.70
50	0.88	0.75	0.67	0.73

Silhouette score measures the quality of clustering by evaluating how similar points within the same cluster are compared to points in other clusters. Higher values indicate better separation between clusters. The proposed method outperforms the existing methods, with the silhouette score improving from 0.74 at 10 epochs to 0.88 at 50 epochs. K-Means and Gaussian Mixture show moderate improvements, while DBSCAN exhibits slower convergence due to its sensitivity to noise and irregular cluster shapes. The CNN-based feature extraction combined with hierarchical clustering enhances the clustering boundary definition, leading to superior performance.

Table 9: Davies–Bouldin Index Comparison

Epochs	Proposed Method	K-Means	DBSCAN	Gaussian Mixture
10	0.42	0.58	0.64	0.55
20	0.38	0.54	0.60	0.50
30	0.35	0.51	0.58	0.47
40	0.32	0.48	0.56	0.45
50	0.28	0.45	0.53	0.42

The Davies–Bouldin index measures the average similarity between clusters, with lower values indicating better clustering quality. The proposed method achieves the lowest value of **0.28** at 50 epochs, demonstrating tighter and more distinct clusters. K-Means and Gaussian Mixture show moderate improvements, while DBSCAN's index remains relatively high due to its sensitivity to noise and non-uniform cluster shapes. The proposed CNN-based hierarchical clustering reduces inter-cluster similarity through improved feature representation and effective distance-based grouping.

Table 10: Clustering Accuracy Comparison

Epochs	Proposed Method	K-Means	DBSCAN	Gaussian Mixture
10	82.4%	74.1%	68.5%	70.2%
20	85.7%	76.3%	70.4%	72.8%
30	89.2%	78.6%	72.7%	75.3%
40	91.4%	80.8%	74.6%	77.6%
50	93.6%	83.2%	76.8%	79.8%

Clustering accuracy measures how well the clustering results match the ground truth labels. The proposed method achieves the highest accuracy of 93.6% at 50 epochs, outperforming K-Means, DBSCAN, and Gaussian Mixture. The CNN-based feature extraction combined with hierarchical clustering enhances cluster cohesion and boundary separation, leading to better accuracy. K-Means and Gaussian Mixture show moderate improvements, while DBSCAN struggles with noisy data. The continuous improvement in clustering accuracy highlights the effectiveness of the combined feature extraction and clustering approach.

CONCLUSION

The proposed CNN-based hierarchical clustering method demonstrates superior clustering performance for genome-wide promoter analysis. Experimental results show that the proposed method achieves a silhouette score of 0.88, a Davies–Bouldin index of 0.28, and a clustering accuracy of 93.6% after 50 epochs. Compared to existing methods such as K-Means, DBSCAN, and Gaussian Mixture, the proposed method exhibits consistent improvement across all evaluation metrics. The CNN-based feature extraction enhances the representational

quality of the promoter sequences, allowing for better feature differentiation and reduced intra-cluster variance. Hierarchical clustering further refines the clustering process by grouping similar sequences based on cosine similarity, ensuring tighter and more distinct clusters. The combination of dimensionality reduction through PCA and max pooling minimizes computational complexity while preserving essential information. The proposed model's ability to handle high-dimensional data and produce biologically meaningful clusters makes it highly suitable for complex genomic analysis. The improved clustering accuracy and separation between clusters demonstrate the robustness and efficiency of the proposed method in identifying promoter subtypes and regulatory patterns.

REFERENCES

[1] Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., ... & Carninci, P. (2009). Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome research*, *19*(2), 255-265.

[2] Manikandan, R., Sara, S. B. V. J., Yuvaraj, N., Chaturvedi, A., Priscila, S. S., & Ramkumar, M. (2022, May). Sequential pattern mining on chemical bonding database in the bioinformatics field. In *AIP Conference Proceedings* (Vol. 2393, No. 1). AIP Publishing.

[3] Kumari SV, Lucas BR, Anitha C, Sangeetha SB, Santhi P, Raja RA, Yuvaraj N. Dense Residual Network-Powered Early Detection of Cardiovascular Diseases Using Multimodal Medical Imaging. J Neonatal Surg [Internet]. 2025Mar.17 [cited 2025Mar.19];14(6S):175-86.

[4] S Chooralil V, Paul V, Megelin Star AA, Kumar RJR, Prabhakaran PN, N. Yuvaraj, Raja RA. Deep Learning-Driven Smart Wearable for Early Prediction and Prevention of Diabetic Complications. J Neonatal Surg [Internet]. 2025Mar.17 [cited 2025Mar.19];14(6S):228-3

[5] Patil, S. C., Madasu, S., Rolla, K. J., Gupta, K., & Yuvaraj, N. (2024, June). Examining the Potential of Machine Learning in Reducing Prescription Drug Costs. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[6] Cherkottu SB, Vijayan S, Mahil J, Chembukkavu J, Gnanaprakash V, Arshath Raja R, Yuvaraj N. AI-Driven Ensemble Deep Learning Framework for Automated Neurological Disorder Diagnosis from MRI Scans. J Neonatal Surg [Internet]. 2025Mar.17 [cited 2025Mar.19];14(6S):187-9.

[7] Wang, Y., Hou, Z., Yang, Y., Wong, K. C., & Li, X. (2022). Genome-wide identification and characterization of DNA enhancers with a stacked multivariate fusion framework. *PLoS Computational Biology*, *18*(12), e1010779.

[8] Lee, D., Yang, J., & Kim, S. (2022). Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nature Communications*, *13*(1), 6678.

[9] Kraft, T., Grützmann, K., Meinhardt, M., Meier, F., Westphal, D., & Seifert, M. (2023). Patient-specific identification of genome-wide DNA-methylation differences between intracranial and extracranial melanoma metastases. *Scientific Reports*, *13*(1), 444.

[10] Liao, X., Lin, R., Zhang, Z., Tian, D., Liu, Z., Chen, S., ... & Su, M. (2024). Genome-wide DNA methylation and transcriptomic patterns of precancerous gastric cardia lesions. *JNCI: Journal of the National Cancer Institute*, *116*(5), 681-693.

[11] Shukla, V., Wang, H., Varticovski, L., Baek, S., Wang, R., Wu, X., ... & Schrump, D. S. (2024). Genome-wide analysis identifies nuclear factor 1C as a novel transcription factor and potential therapeutic target in SCLC. *Journal of Thoracic Oncology*, *19*(8), 1201-1217.

[12] Shukla, V., Wang, H., Varticovski, L., Baek, S., Wang, R., Wu, X., ... & Schrump, D. S. (2024). Genome-wide analysis identifies nuclear factor 1C as a novel transcription factor and potential therapeutic target in SCLC. *Journal of Thoracic Oncology*, *19*(8), 1201-1217.

[13] Li, L., Fei, X., Wang, H., Chen, S., Xu, X., Ke, H., ... & Li, N. (2024). Genome-wide DNA methylation profiling reveals a novel hypermethylated biomarker PRKCB in gastric cancer. *Scientific Reports*, *14*(1), 26605.