Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090108
ISSN 2371-8838 Vol. 9 Num. 1

Inference of Gene Regulatory Networks Based on Heterogeneous Graph Neural Networks

Liming Liu*

School of Information Science and Technology, Yunnan Normal University, Kunming, China
qinghuac1@qq.com
**Corresponding author*

Keywords: Gene Regulatory Network, Heterogeneous Graph, Graph Convolutional Networks

Abstract: Gene Regulatory Networks (GRNs) are central to understanding the mechanisms of gene expression regulation, yet their construction is challenged by node heterogeneity and complex regulatory relationships. Traditional methods often simplify GRNs into homogeneous graphs, overlooking the functional differences between genes and regulatory factors. To address this limitation, we propose a novel GRN construction method, HGRN, based on Heterogeneous Graph Convolutional Networks. By modeling GRNs as heterogeneous graphs comprising two types of nodes—genes and regulatory factors—along with multiple regulatory relationships, and incorporating a multi-channel graph convolution mechanism, our model can separately learn gene expression features and regulatory factor functional features while capturing high-order regulatory dependencies. Experiments on non-specific ChIP-seq datasets demonstrate that this approach outperforms traditional methods in predicting regulatory relationships, significantly improving the accuracy of GRN construction. This study provides a new perspective for the precise inference of gene regulatory networks and offers a powerful tool for elucidating disease mechanisms and predicting drug targets in biomedical research.

1. Introduction

The Gene Regulatory Network (GRN) is a complex system composed of molecules such as transcription factors (TFs) and target genes (TGs), along with their interactions [1]. Its core function is to coordinate biological processes within cells by regulating gene expression. GRNs precisely control the spatiotemporal expression patterns of genes through the binding of transcription factors to DNA or other gene products, thereby regulating critical biological processes such as cell differentiation, metabolism, and stress responses. This multi-layered regulatory mechanism enables cells to adapt to changes in internal and external environments, maintaining homeostasis in biological activities [2]. Research on GRNs not only deepens our understanding of the principles of life regulation but also holds significant value in applications such as drug target discovery, biomarker screening, and the design of personalized therapeutic strategies.

Algorithms for constructing GRNs can be broadly categorized into supervised and unsupervised

methods. Supervised methods leverage known gene-regulator pairs as labels to train models for predicting novel regulatory relationships, with common approaches including Support Vector Machines (SVM) and neural networks. For instance, Mao et al. proposed GNNLink, a novel framework based on Graph Neural Networks (GNNs), for predicting GRNs from single-cell RNA sequencing (scRNA-seq) data [3]. GNNLink first preprocesses scRNA-seq data and then employs Graph Convolutional Networks (GCNs) as an interaction graph encoder to extract gene features and model dependencies between nodes. By transforming the GRN inference problem into a graph link prediction task using matrix completion, GNNLink predicts potential regulatory relationships between genes based on known GRNs and scRNA-seq data.

In contrast, unsupervised methods do not rely on prior knowledge and primarily infer regulatory relationships by analyzing the statistical properties or topological structures of gene expression data. For example, GENIE3 [4], a regression-based method using random forests, was initially developed for bulk RNA sequencing data but has also been applied to scRNA-seq data. GENIE3 constructs multiple regression trees to predict the expression levels of each gene and infers regulatory relationships based on the importance of regulators in these predictions. Both approaches have their strengths: supervised methods offer higher accuracy but depend on labeled data, while unsupervised methods are more suitable for exploratory analysis of large-scale datasets.

Traditional methods often simplify GRNs into homogeneous graphs, assuming that all nodes (genes, transcription factors) share identical attributes and functions, thereby overlooking their inherent differences in biological roles and regulatory mechanisms. This simplification limits the model's ability to represent complex regulatory relationships, particularly in distinguishing between different types of regulatory interactions such as activation and repression, as well as capturing higher-order topological structures like feedback loops and regulatory modules. To address these limitations, this paper proposes HGRN, a novel GRN construction method based on Heterogeneous Graph Convolutional Networks (HGCN). By modeling GRNs as heterogeneous graphs comprising two types of nodes—genes and regulatory factors—along with multiple regulatory relationships, and incorporating a multi-channel graph convolution mechanism, the model can separately learn gene expression features and regulatory factor functional features while explicitly modeling different types of regulatory interactions. This approach not only overcomes the limitations of traditional homogeneous graph assumptions but also captures long-range regulatory dependencies through high-order neighborhood aggregation, providing a new research direction for the precise inference and functional analysis of GRNs.

2. Materials and methods

2.1. Dataset

In this study, experimental validation was conducted on a non-specific ChIP-seq dataset [5], specifically utilizing the TFs +1000 dataset as the benchmark. This dataset integrates high-throughput experimental data such as ChIP-seq and RNA-seq, providing potential regulatory relationships between genes and transcription factors, making it a commonly used benchmark for evaluating GRN construction methods. Each sample in the dataset includes a gene expression matrix, transcription factor binding site information, and known transcription factor-gene regulatory pairs, offering rich node features and edge relationship information for the construction of heterogeneous graphs.

2.2. The Construction of Heterogeneous Graphs for Gene Regulatory Networks

To more accurately model the complex regulatory relationships within GRN, this study

constructs the GRN as a heterogeneous graph, which includes two types of nodes—TGs and TFs—as well as multiple types of edge relationships. The TF-TG relationship represents the regulatory interaction between a transcription factor and its target gene, while the TF-TF relationship captures interactions between transcription factors, such as protein-protein interactions or cooperative regulatory relationships. Based on these definitions of nodes and edges, the GRN is modeled as a heterogeneous graph G=(V, E, T), where V denotes the set of nodes, encompassing both gene nodes and transcription factor nodes; E represents the set of edges, including TF-TG and TF-TF relationships; and T signifies the set of node types and edge types, which are used to distinguish between different categories of nodes and edges. By constructing the heterogeneous graph, the proposed method explicitly models the heterogeneity between genes and transcription factors, enabling a more nuanced representation of the regulatory network.

2.3. Heterogeneous Graph Convolutional Network

HGRN is a deep learning model based on a multi-channel heterogeneous graph convolutional network, designed to learn feature representations of genes and transcription factors within gene regulatory networks. We perform graph convolution operations on different types of edges separately and aggregate the convolution results across various edge types through summation. The core idea of GCN is to update node representations by aggregating features from a node and its neighbors. For a given regulatory relationship, the node feature update formula for the l-th layer of GCN is as follows:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{\hat{d}_i \hat{d}_j}} W^{(l)} h_j^{(l)} \right) \tag{1}$$

Here, $h_j^{(l)}$ represents the feature representation of node i at the l-th layer, $N(i)$ denotes the set of neighbors of node i, $\hat{d}_i$ and $\hat{d}_j$ are the normalized degree coefficients of nodes i and j, respectively, $W^{(l)}$ is a learnable weight matrix, and σ is a non-linear activation function.

3. Results

To comprehensively evaluate the classification performance of the model, this study employs five widely used evaluation metrics: Accuracy, Precision, F1-score, Area Under the ROC Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). Among these, Accuracy measures the proportion of correctly predicted samples out of the total samples and is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. However, in datasets with class imbalance, accuracy may not fully capture the model's performance. Therefore, this study further introduces Precision and Recall: Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive, and its calculation formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

To provide a balanced assessment of both Precision and Recall, this study employs the F1-score, which is calculated as follows:

$$\text{F1 Score} = \frac{2TP}{2TP + FN + FP} \tag{4}$$

The F1-score is particularly important in scenarios with imbalanced class distributions, as it effectively balances the trade-off between Precision and Recall.

Additionally, this study utilizes AUROC and AUPRC for threshold-independent performance evaluation. AUROC reflects the model's ability to distinguish between positive and negative samples across different classification thresholds, with values closer to 1 indicating better classification performance. However, in cases of severe class imbalance, the PR curve provides more meaningful insights than the ROC curve. Therefore, this study further incorporates AUPRC. By comprehensively considering the variations in Precision and Recall, AUPRC offers a more accurate reflection of the model's performance when positive samples are scarce.

In this study, 5-fold cross-validation was employed to evaluate the model's performance. Specifically, the dataset was randomly divided into 5 subsets of equal size. In each iteration, one subset was used as the validation set, while the remaining 4 subsets served as the training set. This process was repeated 5 times, and the average results from these iterations were taken as the final performance metrics. This approach not only maximizes the utilization of limited data but also minimizes evaluation bias caused by data partitioning variations. During the training process, the number of epochs was set to 100 to ensure the model fully learned the data features. The results on the non-specific dataset are presented in Table 1.

Table 1: Results on the Non-Specific ChIP-seq Dataset

	AUROC	Accuracy	Precision	F1	AUPRC
hESC	0.8305	0.8305	0.7469	0.8551	0.8735
hHEP	0.8382	0.8391	0.7592	0.8608	0.8777
mDC	0.9731	0.9285	0.9257	0.9279	0.9664
mESC	0.9541	0.9096	0.879	0.9133	0.9357
mHSC-E	0.8998	0.8998	0.8334	0.9091	0.9167
mHSC-GM	0.9266	0.9267	0.8725	0.9318	0.9361
mHSC-L	0.9587	0.9589	0.9244	0.9606	0.9619

Table 2: Comparison of HGRN with Other Models

	hESC	hHEP	mDC	mESC	mHSC-E	mHSC-GM	mHSC-L
HGRN	0.83	0.83	0.97	0.95	0.89	0.92	0.95
GENELink	0.69	0.70	0.78	0.77	0.75	0.77	0.67
GNE	0.63	0.62	0.67	0.69	0.56	0.62	0.59
DeepSEM	0.53	0.55	0.56	0.56	0.59	0.61	0.60
PCC	0.51	0.53	0.49	0.57	0.59	0.66	0.62
MI	0.48	0.46	0.48	0.54	0.64	0.74	0.65
SCODE	0.51	0.51	0.49	0.52	0.55	0.57	0.56
GRNBOOST2	0.50	0.49	0.52	0.55	0.64	0.70	0.64
GENIE3	0.48	0.47	0.52	0.56	0.63	0.70	0.65

To comprehensively evaluate the performance of our model, we directly compared our results with the method proposed by Chen et al [6]. The detailed results are presented in Table 2. The findings demonstrate that HGRN outperforms other models across all datasets. The strength of HGRN lies in its ability to effectively learn features under different regulatory types through heterogeneous graph convolutional networks. By leveraging the heterogeneous graph structure, the method can more thoroughly extract features from diverse regulatory relationships, playing a pivotal role in enhancing prediction accuracy.

4. Conclusion

This study proposes HGRN, a GRN inference method based on Heterogeneous Graph Convolutional Networks (HGCN). By modeling GRNs as heterogeneous graphs comprising diverse node types and multiple edge types, our approach effectively captures the inherent heterogeneity and complex interactions within biological systems. The multi-channel graph convolution mechanism enables the model to learn node-specific features and higher-order dependencies, overcoming the limitations of traditional homogeneous graph-based methods. Experimental results on non-specific ChIP-seq datasets demonstrate the superior performance of this method in predicting regulatory relationships, highlighting its potential for GRN inference. Future research could explore integrating additional biological prior knowledge and applying this framework to construct disease-specific GRNs, thereby providing deeper insights into gene regulation and its applications in biomedical research.

References

[1] Xu J, Zhang A, Liu F, et al. STGRNS: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data [J]. Bioinformatics, 2023, 39(4): btad165.
[2] McCall M N. Estimation of gene regulatory networks [J]. Postdoc journal: a journal of postdoctoral research and postdoctoral affairs, 2013, 1(1): 60.
[3] Mao G, Pang Z, Zuo K, et al. Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks [J]. Briefings in Bioinformatics, 2023, 24(6): bbad414.
[4] Huynh-Thu V A, Irrthum A, Wehenkel L, et al. Inferring regulatory networks from expression data using tree-based methods[J]. PloS one, 2010, 5(9): e12776.
[5] Garcia-Alonso L, Holland C H, Ibrahim M M, et al. Benchmark and integration of resources for the estimation of human transcription factor activities [J]. Genome research, 2019, 29(8): 1363-1375.
[6] Chen G, Liu Z P. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data [J]. Bioinformatics, 2022, 38(19): 4522-4529.