Gene regulatory network reconstruction based on two-layer neighbor overlapping perceptual graph convolution network

Wenxuan Luo[a], Zhiqiong Wang[a*], Xinyang Li[a], Yameng Guo[a], Jiaxin Cao[b], Yifan Feng[a]
[a]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110819, China; [b]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[*] Corresponding author: wangzq@bmie.neu.edu.cn

ABSTRACT

Reconstructing gene regulatory networks (GRNs) is a fundamental challenge in bioinformatics that aims to unravel the complex relationships between genes and their regulators. Graph convolutional neural networks have shown more significant improvements in this field than traditional methods. However, GCNs rely heavily on smooth node features rather than graph structures. To address this limitation, Two-layer Neighbor Overlapping Perceptual Graph Convolution Network (Tnop-GCN) is proposed, that jointly learns local and global structural features by PageRank and DeepWalk. Experiments on DREAM4 dataset demonstrate that Tnop-GCN outperforms many other gene regulatory network reconstruction methods.

Keywords: Gene regulatory networks, graph convolutional neural networks, link prediction, PageRank, DeepWalk.

1. INTRODUCTION

The processes of gene expression are shown to be mutually regulated rather than to be in isolation. And such complex regulatory relationships construct gene regulatory networks (GRNs), which represent the molecular dynamic processes of many organisms including human beings [1]. As the accurate reconstruction of GRNs is of significance for precision medicine and many related clinical applications, the study of gene regulatory networks holds significant importance for propelling research in the life sciences, uncovering the mechanisms of diseases, and fostering the development of biotechnological advancements [2]. Traditional methods for reconstructing GRNs include model-free and model-based approaches. Model-free reconstruction methods do not rely on any preset models and directly identify regulatory relationships from gene expression data. Model-based methods involve constructing mathematical models to describe the dynamic relationships between genes and learning the parameters of these models. Common models include Boolean network models[3][4], and Bayesian network models[5][6].

Graph representation learning, also known as network embedding, aims to map the vertices of a graph to a low-dimensional vector space while preserving as much of the vertices' topological structure as possible [7]. These vector representations support various network analysis tasks, such as node classification, link prediction, and community detection. Recently, the efficiency of graph representation learning has been significantly improved by using deep learning algorithms like Skip-gram[8] and convolutional networks [9]. For example, DeepWalk [10] and Node2vec[8] algorithms generate node sequences through random walks and then employ methods similar to those used in natural language processing to learn vector representations of nodes. Besides, PageRank algorithm is a web ranking algorithm based on link analysis which calculates the importance and ranking of web pages by analyzing the link relationship between web pages, and then deduces the quality and influence of web pages. The PageRank algorithm can treat the entire Internet as a huge directed graph, and determine the relevance and ranking position of the webpage by iterating the PageRank value of each webpage, so it is usually used for both webpage ranking and graph representation.

With advancements in deep learning technology, novel methods tend to reconstruct GRNs by performing link prediction task, with the aim for predicting potential connections using node features and an incomplete prior network. However, the out-degree of GRNs follows a power-law distribution and the in-degree follows an exponential distribution, making it difficult for many link prediction algorithms to capture this unique structural feature. To overcome the limitation, in this study, a novel GRN reconstruction inference based on GCN and multi-feature from PageRank and DeepWalk is proposed to explore the potential regulatory relationships from gene expression data. To fully leverage the information multi-

International Conference on Future of Medicine and Biological Information Engineering (MBIE 2024), edited by Yudong Yao, Xiaoou Li, Xia Yu, Proc. of SPIE Vol. 13270, 132701A · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3040808

dimensional data, GCN is employed to extract node features first, yielding meaningful node feature matrix. Then, PageRank and DeepWalk are applied to capture the global network structural information and the local network feature information, respectively. Lastly, three types of features are added by weight, thereby achieving link prediction for GRN reconstruction. Experiments are carried out on DREAM4 multifactor datasets with different scale, which shows the proposed method outperforms many other baseline methods.

2. METHOD

2.1 Preliminaries

Notations. For a given undirected network $G=(V,E)$, where V contains all N nodes in the network, $v_i \in V$, E stands for the edge between nodes $(v_i,v_j) \in E$, The elements of the adjacency matrix $A \in \{0,1\}^{N\times N}$ is binary, $A_{ij}=1$ iff $(v_i,v_j) \in E$. Degree matrix $D_{ii}=\sum_j A_{ij}$; The feature matrix of the node is $X \in R^{N\times C}$, where N is the number of nodes and C is the dimension of the feature.

A graph convolution neural network for link prediction. Given the graph G and the feature X, the graph convolution neural network learns meaningful node representations by iteratively aggregating the transformation representations of neighbor nodes in each *i* th GCN layer, as shown below,

$$H^{(l+1)}=\sigma(\tilde{A}_{GCN}H^{(l)}W^{(l)}) \tag{1}$$

where $\tilde{A}_{GCN} \in R^{N\times N}$ is an adjacency matrix normalized in different ways according to each GCN architecture, $W^{(l)} \in R^{d^{(l)}\times d^{(l+1)}}$ is a trainable weight matrix, and $H^{(0)}$ is a node feature matrix $X \in R^{N\times F}$. After stacking L GCN layers, use the node representation $H^{(L)}$ to predict the existence of each link (i,j):

$$\hat{y}_{ij}=\sigma(s(h_i^{(L)},h_j^{(L)})) \tag{2}$$

where $s(\cdot,\cdot)$ is a function, $e.g.$, inner product or MLP, and $h_i^{(L)}$ is the representation of the node i from $H^{(L)}$.

Deepwalk. For each point $u \in V$,define $N_S(u) \subset V$ as a network neighbor that generates a node u through a neighbor sampling strategy S. The goal of Node2vec is: with a given vertex U maximize its neighbors $N_S(u) \subset V$ and get its logarithmic likelihood function in low-dimensional space. That is:

$$D(u)=\max_f \sum_{u\in V} \log \Pr(N_S(u) \mid f(u)) \tag{3}$$

Then, the node representation matrix L is gotten, which contains the low-dimensional representation of each node.

Pagerank. The PageRank formula is shown as follow,

$$PR(a)_{i+1}=\sum_{i=0}^{n} \frac{PR(Ti)_i}{L(Ti)} \tag{4}$$

$PR(Ti)_i$:PR value of other nodes (pointing to node a); $L(Ti)$:the number of outgoing links of other nodes (pointing to node a); i :cycle number.

2.2 Two-layer neighbor overlapping perceptual graph convolution network

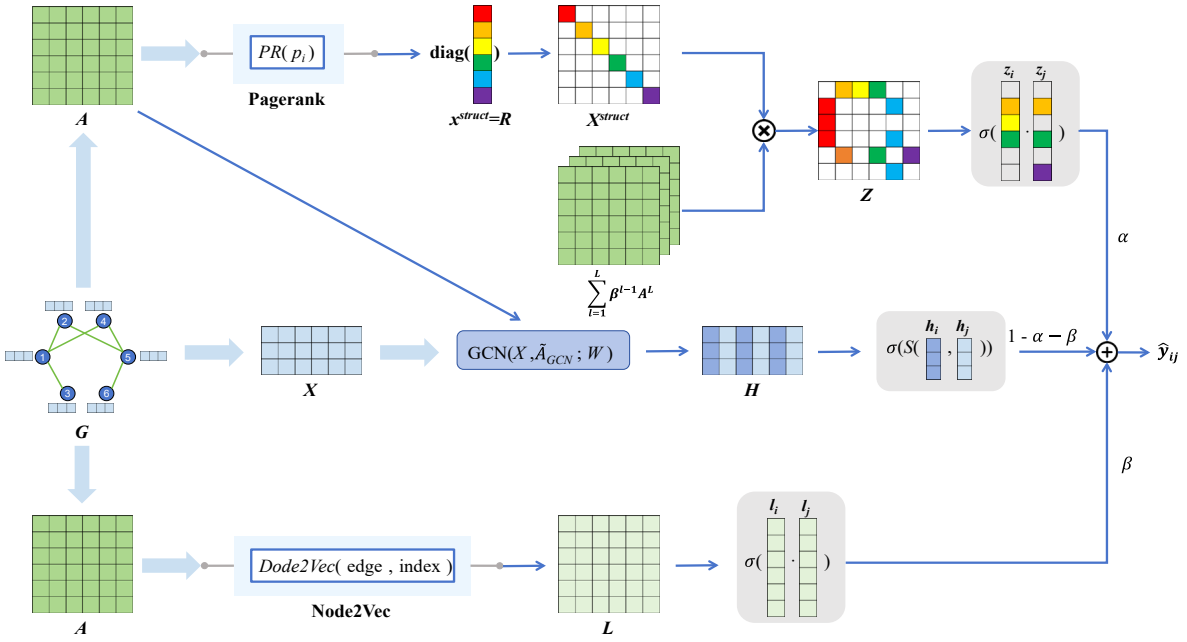

Figure 1: The Tnop-GCN framework for link prediction. Tnop-GCN learns structural features from an adjacency matrix and estimates similarity scores based on overlapped neighborhoods. Tnop-GCN first uses Node2Vec and Pagerank to generate local structural matrix and global structural matrix. Then Tnop-GCN calculates both similarity scores from each representation matrix L, Z and H and computes the convex combination of three scores by a trainable parameter α.

Link prediction based on local structural feature: As the low-dimensional node representation matrix L is gotten, each i th row vector of L_i involves all the feature of the adjacent nodes of node i respectively. The existence score of a link between node i and node j is:

$$y_{ij} = D(i)D(j)^T \tag{5}$$

Link prediction based on global structural feature: We use structural feature vectors $PR(a) \in R_{N\times 1}$ to construct diagonal matrix $X^{struct} \in R_{N\times N}$ to maintain the respective feature of each node after aggregation,

$$X^{struct} = diag\left(PR(a)\right) \tag{6}$$

Then, to consider the number of overlapping neighbors, the Tnop-GCN aggregate the feature of the neighbors by multiplying the non-normalized adjacency matrix A and get the similar node information matrix Z :

$$Z = AX^{struct} \tag{7}$$

Further, in order to consider the neighbors with multi-hop overlap, the multi-hop settings will be extended as:

$$Z = g_\Phi\left(\sum_{l=1}^{L} \beta^{l-1} A^l X^{\text{struct}}\right) \tag{8}$$

where β is a super parameter, controlling the weight of near and far neighbors, g_ϕ is MLP, that controls the scale of L.

With the similar node information matrix Z, each *i*-th row vector of Z involves all the feature of the neighboring nodes of node *I*. Then we can get the existence score of the link between node i and node j:

$$z_i^T z_j = \sum_{k \in N(i) \cap N(j)} \left(x_k^{struct} \right)^2 \tag{9}$$

Domain overlapping perceptual aggregation scheme: As is shown in Figure 1, with a link (i, j), Tnop-GCN calculates both similarity scores from each representation matrix *L*, *Z* and *H* and computes the convex combination of three scores by a trainable parameter α as follows:

$$\hat{y}_{ij} = \alpha \cdot \sigma\left(z_i^T z_j \right) + \beta \cdot \sigma\left(l_i^T l_j \right) + (1 - \alpha - \beta) \cdot \sigma\left(s\left(h_i, h_j \right) \right) \tag{10}$$

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Dataset and baseline methods

Dataset. The experiments are carried out on DREAM4 multifactorial dataset which comprises two scales of networks: small-scale networks containing 10 genes and large-scale networks containing 100 genes. Each scale is represented by five independent datasets. These networks possess diverse topological structures, selected to simulate the biological systems of Escherichia coli or Saccharomyces cerevisiae. For the networks of size 10, each network is constituted by five time series, whereas for those of size 100, each network is constituted by ten time series. Each time series encompasses 21 data points.

Baselines. To demonstrate the effectiveness of Tnop-GCN, the performance of Tnop-GCN is compared with four GRN reconstruction model, DeepSEM [11], GENIE3 [12], GNNLink [13] and ARACNe-ap [14]. DeepSEM is a neural network version of the structural equation model (SEM), which explicitly models the regulatory relationships among genes. GENIE3 is a model based on feature selection with tree-based ensemble methods and is the best performer in the DREAM4 in silico multifactorial challenge. GNNLink leverages known GRNs to deduce the potential regulatory interdependencies between genes and shows great robustness and accuracy. ARACNe-ap can build complex regulatory networks from hundreds of gene expression profiles, with an Adaptive Partitioning strategy (AP) for estimating the mutual information.

3.2 Performance on 10-Gene Networks

As is shown in Table 1, the comparison involved Tnop-GCN, GENIE3, and deepSEM. Tnop-GCN exhibited notably superior average AUC values compared to GENIE3 and deepSEM. It is shown that GENIE3 is a little better than Tnop-GCN in the range of Net2. This may be attributed to the merit in scenarios where networks exhibit specific characteristics that align with its inference methodology. However, the performance of Tnop-GCN is superior on all of the networks except Net2. Despite the exceptional performance of GENIE3 in Net2, Tnop-GCN's overall superiority in average AUC values across other sub-networks underscores its pivotal role and significance in gene network inference algorithms. This consistency reaffirms the importance and necessity of Tnop-GCN in achieving reliable and accurate gene regulatory network predictions.

Table 1: The performance comparison on DREAM4 10-gene Networks (Using AUC).

	DeepSEM	GENIE3	Tnop-GCN
Net1	0.539	0.669	**0.833**
Net2	0.598	**0.711**	0.667
Net3	0.627	0.644	**0.667**
Net4	0.700	0.378	**0.750**
Net5	0.574	0.691	**0.750**
Avg	0.608	0.619	**0.733**

3.3 Performance on 100-Gene Networks

As is shown in Table 2, Tnop-GCN demonstrated the highest average AUC (0.670) among the algorithms tested on 100-Gene Networks, confirming its superior performance and essential role in gene network analysis. It has been illustrated that ARACNe-ap outperforms Tnop-GCN marginally when dealing with Net3 and Net4. This superiority of ARACNe-ap could be attributed to its specialized adaptability to the unique structural features of these specific networks, where its algorithmic approach might exploit network intricacies more effectively. Conversely, in other networks, Tnop-GCN consistently surpassed its counterparts with significantly higher average AUC values. This overall trend highlights the robustness and general applicability of Tnop-GCN across diverse network architectures.

Table 2: The performance comparison on DREAM4 100-gene Networks (Using AUC).

	GNNLink	ARACNe-ap	deepSEM	Tnop-GCN
Net1	0.537	0.602	0.550	**0.757**
Net2	0.646	0.568	0.535	**0.674**
Net3	0.509	**0.655**	0.530	0.641
Net4	0.576	**0.645**	0.510	0.631
Net5	0.576	0.627	0.525	**0.645**
Avg	0.569	0.619	0.530	**0.670**

3.4 Discussion

In summary, our study underscores Tnop-GCN as an exceptional algorithm for gene network inference, consistently achieving superior performance across most sub-networks in both 10-gene and 100-gene networks. While some algorithm occasionally outperformed Tnop-GCN in specific network contexts, the overall effectiveness of Tnop-GCN in producing higher average AUC values reaffirms its critical role in advancing gene network analysis methodologies. Future research endeavors should continue to explore the nuanced interactions between algorithmic design and network topology to enhance the precision and applicability of gene regulatory network inference techniques in biological research.

4. CONCLUSION

The research introduces a graph convolution neural network (Tnop-GCN) based on two-layer neighbor overlap, which can extract local and global structure information, which is the key element of link prediction. Tnop-GCN learns useful structural features from the adjacency matrix and estimates overlapping neighbors for link prediction. The research also adaptively combines Tnop-GCN and feature-based Tnop-GCN to consider structural features and input node features. In addition, the research also evaluates Tnop-GCN to prove its efficiency. In the future work, the plan is to further develop Tnop-GCN to promote more heuristic methods based on link prediction and improve scalability through efficient sparse matrix computation.

ACKNOWLEDGEMENT

The work is supported by National Training Program of Innovation and Entrepreneurship for Undergraduates (240240), National Natural Science Foundation of China (62072089) and the Fundamental Research Funds for Central Universities (N2424010-19).

REFERENCES

[1] Etienne F and Nathan M, "Gene regulatory network investigation using ordinary differential equations", Methods Mol. Biol., (2021).

[2] Mingkun F, Xiangtian J, "Inference of gene regulatory networks based on nonlinear ordinary differential equations", Bioinformatics, (2020).
[3] Ning S, Zexuan Z, Ke T, David P, Shan H, "ATEN: And/or tree ensemble for inferring accurate boolean network topology and dynamics", Bioinformatics, (2020).
[4] Shohag B, Yung-Keun K, "A boolean network inference from time-series gene expression data using a genetic algorithm", Bioinformatics, (2018).
[5] Polina S, Jack K, Niko B, "Discovering gene regulatory networks of multiple phenotypic groups using dynamic bayesian networks", Brief. Bioinf (2022).
[6] Hamda B. A, Michael G. M, "Dynamic bayesian network learning to infer sparse models from time series gene expression data", IEEE/ACM Trans. Comput. Biol. Bioinf., (2022).
[7] Seongjun Y et al. “Neo-GNNs: Neighborhood Overlap-aware Graph Neural Networks for Link Prediction” Neural Information Processing Systems (2022).
[8] Aditya G, Jure L, "node2vec: Scalable Feature Learning for Networks", KDD: proceedings. International Conference on Knowledge Discovery & Data Mining, (2016).
[9] Thomas N. K, Max W, "Semi-Supervised Classification with Graph Convolutional Networks", CoRR, (2016).
[10] Bryan P, Rami A, Steven S, "Deepwalk: Online learning of social representations", in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, (2014).
[11] Shu, H., Zhou, J., Lian, Q. et al, "Modeling gene regulatory networks using neural network architectures", Nat Comput Sci 1, (2021).
[12] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods", PLoS ONE 5(9): e12776 (2010).
[13] Mao G, Pang Z, Zuo K, Wang Q, Pei X, Chen X, Liu J, "Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks", Brief Bioinform (2023).
[14] Lachmann A, Giorgi FM, Lopez G, Califano A, "ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information", Bioinformatics, (2016).