ISSN:2790-1688 Volume-14-(2025)

Construction of an Exosome Antioxidant Pathway Recognition Algorithm by Integrating Bioinformatics Mining with Graph Neural Networks

Chuntian Liu

State Key Laboratory of Biobased Material and Green Papermaking, School of Bioengineering, Qilu University of Technology, Shandong Academy of Sciences, Jinan, 250353, China

1574643562@qq.com

Abstract. It is very important to identify and understand the antioxidant pathways in exosomes efficiently, so as to deeply explore the biological functions of exosomes and develop related therapeutic methods. This article aims to construct an exosome antioxidant pathway identification algorithm that combines bioinformatics mining and graph neural network (GNN). In this article, the characteristics of gene expression, protein interaction and metabolic pathway are extracted, and GNN training is optimized by random node sampling and random walk sampling. Then, an efficient and accurate prediction model is proposed by combining the improved graph convolution network (GCN) and multi-task convolutional neural network (CNN). The results show that the accuracy of the proposed algorithm on the test set is 89.7%, and the F1 score is 88.7%, which is obviously superior to the traditional method. In addition, the model successfully identified several key molecules, such as SOD1 and GPX1, and their functional annotations were highly consistent with the existing literature. These findings can verify the effectiveness of the algorithm and provide a new perspective for understanding the biological function of exosomes in antioxidant mechanism.

Keywords: Exocrine body; Antioxidant pathway; Graph neural network; Node sampling; Multiomics data.

1. Introduction

Exosomes are tiny vesicles secreted by cells, which play an important role in many physiological and pathological processes besides participating in intercellular information transmission. Antioxidant pathway is one of the key mechanisms to maintain the stability of intracellular environment, which is of great significance in antioxidant stress and disease prevention [1]. However, it is very important to effectively identify and understand the antioxidant pathways in exosomes for further exploring the biological functions of exosomes and developing related treatment methods [2]. In recent years, the development of bioinformatics provides a powerful tool for life science research. By analyzing large-scale biological data, it can reveal the laws of complex biological processes such as gene expression patterns and protein interaction networks [3]. However, traditional bioinformatics mining methods are faced with problems such as high data dimension, insufficient sample size and difficult feature extraction [4]. These methods are even more inadequate when dealing with such a complex biological system as exosomes.

As a new deep learning model, GNN has great potential in processing structured data. GNN operates directly on the graph structure, which can effectively capture the relationship between nodes and their roles in the network, and is suitable for processing data with complex topological structure such as biomolecular networks [5]. Although GNN has made remarkable achievements in the fields of drug discovery and disease prediction, there are few attempts to apply it to exosomes research, especially to identify antioxidant pathways. Based on this background, this study attempts to combine bioinformatics mining with GNN to break through the bottleneck of existing technology and promote the development of exosome antioxidant pathway identification technology.

The data related to exosomes come from a wide range of sources, including public databases and laboratory experimental results. These data often contain a lot of noise and redundant information, so how to clean and integrate them efficiently and make them suitable for subsequent analysis needs to be solved urgently. On this basis, the advanced bioinformatics mining technology is used to

preliminarily analyze the data and determine the candidate molecules that may be related to the antioxidant pathway. Then, design an appropriate GNN model, map the candidate molecules and their interactions into the graph structure, and use GNN's powerful characterization ability to analyze them deeply.

2. Theoretical basis

2.1 Biological information mining technology

Bioinformatics is an interdisciplinary field, which uses computational methods to understand and process biological data. With the development of high-throughput sequencing technology and large-scale genome project, it has become an indispensable part of modern biological research. Data preprocessing is the basic step of all data analysis. Exosomes have a wide range of sources and forms, and often contain a lot of noise and redundant information [6]. Therefore, it is extremely important to clean and integrate these data efficiently. Common data preprocessing methods include removing low-quality sequences, standardizing expression levels, and properly normalizing [7].

Cluster analysis is an effective unsupervised learning method, which can divide samples into different groups according to their similarity. For example, when analyzing the gene expression profile of exosomes, hierarchical clustering or K-means clustering algorithm can be used to identify co-expression patterns, and then potential functional modules can be found [8]. Differential expression analysis is also an important method to identify genes or proteins that have changed significantly under different experimental conditions. Path enrichment analysis is also one of the commonly used tools in bioinformatics. Mapping the list of genes of interest to known biological paths and evaluating whether these paths are significantly enriched can reveal the key molecular mechanisms involved in specific biological processes.

2.2 GNN principle and its application

GNN is a new deep learning model, which shows great potential in dealing with non-Euclidean structured data. Traditional deep learning models, such as CNN and recurrent neural network(RNN), are mainly suitable for processing regular grid data, such as images and texts [9]. GNN can operate directly on the graph structure, thus effectively capturing the relationship between nodes and the role they play in the network.

The basic idea of GNN is to update the representation of each node by recursively aggregating the information of neighboring nodes. Given a graph $G = (V, E)$, where V represents a node set and E represents an edge set. For each node $v_i \in V$, its initial representation $h_i^{(0)}$ may be the original feature vector or other predefined initialization values. Then, the formula is updated by multi-layer iteration:

$$h_i^{(l+1)} = \sigma\left(W^{(l)} \sum_{j \in N(i)} h_j^{(l)} + b^{(l)}\right) (1)$$

Here, $N(i)$ represents the neighbor node set of node $W^{(l)}$, $b^{(l)}$ and l are the weight matrix and bias term of the σ layer respectively, and FF is the activation function. After several iterations, the final node representation can be used for various downstream tasks, such as node classification and link prediction.

In the field of bioinformatics, GNN has been successfully applied in many scenarios. For example, in the process of drug research and development, researchers use GNN to model the molecular structure to predict the activity and toxicity of compounds [10]. Similarly, in disease prediction, GNN is also used to analyze gene interaction networks and identify mutations that may cause diseases. However, although GNN performs well in many tasks, it also encounters some difficulties in practical application. Problems such as over-fitting and low computational efficiency occur from time to time.

2.3 Exosomes antioxidant pathway

Antioxidant pathway is a series of biochemical reactions to antioxidant stress in organisms. Oxidative stress is a state caused by excessive production of free radicals and other reactive oxygen species (ROS), which will cause cell damage and even death if it is not removed in time [11]. Antioxidants can maintain the stability of the intracellular environment by neutralizing these harmful substances.

As the key medium of intercellular communication, exosomes can not only carry many biological macromolecules such as nucleic acid and protein, but also affect the state and behavior of recipient cells. Studies have shown that exosomes are rich in antioxidant enzymes and antioxidant small molecules, which can help target cells resist oxidative stress [12]. In addition, exosomes may also enhance the antioxidant capacity of cells by regulating intracellular signal transduction pathways.

However, the research on the antioxidant pathway of exosomes is still in the primary stage. On the one hand, exosomes have a complex structure and a variety of components, so it is difficult to fully understand their specific mechanism of action. On the other hand, most of the existing studies focus on a single type of exosomes or specific antioxidant components, lacking a systematic and global perspective. Therefore, it is particularly important to construct an exosome antioxidant pathway identification algorithm that can comprehensively consider various factors.

3. Methodology

3.1 Data collection and pretreatment

In order to ensure the effectiveness of the algorithm, multiple omics data related to exosomes are integrated from multiple public databases. Gene expression data are derived from TCGA(The Cancer Genome Atlas) and GEO(Gene Expression Omnibus) to screen differentially expressed genes. The data of protein interaction network is taken from STRING database, which provides the known information of protein-protein interaction. Metabolic pathway information is obtained through Kegg (Kyoto encyclopedia of genes and genomics) database, which is used to label the functional properties of candidate molecules [13]. These data are rich in sources, including different types of biomolecules and their complex interactions.

In the data preprocessing stage, firstly, the data is cleaned to remove samples with more missing values or poor quality. Then, the gene expression data were standardized, and the influence caused by the difference of experimental conditions was eliminated through normalization operation:

$$x' = \frac{x-\mu}{\sigma} \tag{1}$$

Where x stands for raw data, μ and σ stand for mean and standard deviation respectively. Finally, based on statistical analysis and machine learning methods, feature selection is carried out, and key features related to antioxidant function are screened out.

3.2 Node sampling strategy

Biomolecular networks are usually large in scale and complex in structure, and direct application of GNN may lead to insufficient computing resources. Therefore, two different node sampling strategies are adopted to optimize the training process: Random Node Sampling and Random Walk Sampling.

The core idea of random node sampling is to randomly select some nodes and their first-order neighbors from the whole network to form a subgraph. The advantage of this method is that it can significantly reduce the computational complexity while retaining the key features of the network structure. Randomly select a central node v_i; Extracting the first-order neighbor set $N(v_i)$ of v_i; Combine v_i and $N(v_i)$ into a subgraph $G_s = (V_s, E_s)$. Where V_s is a node set and E_s is an edge set.

The mathematical description of this method is:

$$G_s = \{v_i \cup N(v_i)\},\ \forall v_i \in V \tag{2}$$

In this way, multiple subgraphs can be generated and sent to GNN as input data for training.

Random walk sampling is a more complex sampling method, which aims to capture the long-range dependence in the network. Its basic principle is to select the visited nodes and their connection relationships by simulating the random walk process.

Starting from any node v_0 in the network, the initial state $S_0 = \{v_0\}$ is defined. According to the transition probability $P(v_{t+1}|v_t)$, the next node is visited in turn until the preset number of steps T is reached, and all the nodes visited during the walk and their edges are combined into a subgraph G_w.

The probability transition equation of random walk is:

$$P(v_{t+1}|v_t) = \frac{w(v_t,v_{t+1})}{\sum_{u\in N(v_t)} w(v_t,u)} \tag{3}$$

Where $w(v_t, v_{t+1})$ represents the weight between nodes v_t and v_{t+1}.

With the help of random walk sampling, more nodes can be covered and the deep association hidden in the network can be revealed.

Figure 1 shows the implementation process of these two sampling strategies. Random Node Sampling on the left and Random Walk Sampling on the right. The former pays more attention to the preservation of local structure, while the latter emphasizes the capture of global information.

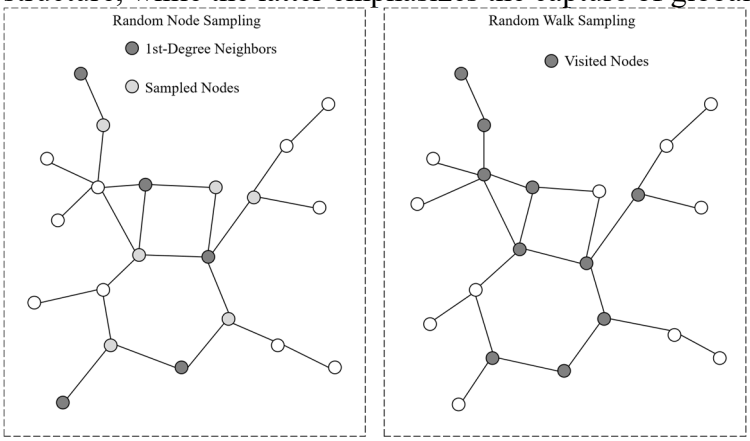

Figure 1 GNN architecture schematic

3.3 GNN architecture design

After completing node sampling, the generated subgraph will be input into GNN for training. This article uses an improved GCN architecture. Its core idea is to update the node representation with the help of message passing mechanism and finally achieve the classification task.

The message passing process of GCN can be formalized as the following equation:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{4}$$

Where $H^{(l)}$ represents the node representation matrix of the l layer; $\tilde{A} = A + I$ is adjacency matrix plus identity matrix; $\widetilde{D}$ is the degree matrix of $\tilde{A}$; $W^{(l)}$ is a learnable weight matrix; σ is the activation function. After many iterations, the representation of each node gradually fuses the information of its neighbors, thus obtaining higher-level semantic features.

In order to further improve the performance of the model, an innovative architecture similar to multi-task CNN is introduced to extract more complex biomolecular network features (see Figure 2). The architecture consists of three core modules: first, the convolution layer extracts local features by sliding window operation; Secondly, the pooling layer reduces the dimension of features to reduce the computational complexity; Thirdly, the fully connected layer maps the extracted high-dimensional features to specific classification labels to complete the end-to-end feature learning and classification tasks.

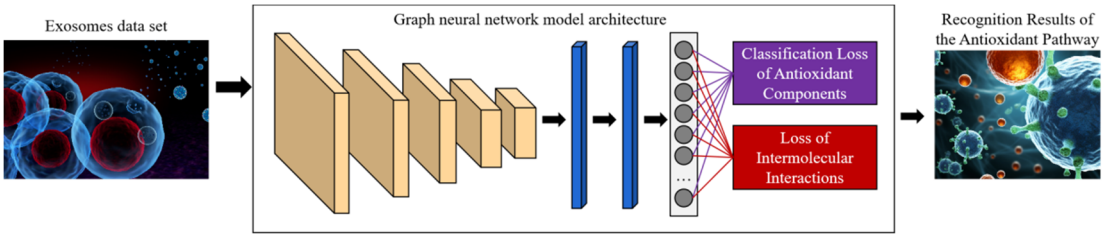

Figure 2 Antioxidant pathway identification model of exosomes

Assuming that the input data is $X \in \mathbb{R}^{n \times d}$, where n is the number of samples and d is the feature dimension, the output of the convolution layer can be expressed as:

$$Y = f(W * X + b \tag{5}$$

Where W is the convolution kernel, b is the bias term, and f is the activation function.

3.4 Model training and optimization

In the model training stage, the cross entropy loss function is used to measure the gap between the prediction results and the real labels:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{6}$$

Where N is the total number of samples, C is the number of categories, and y_{ij} and $\hat{y}_{ij}$ respectively represent the true label and prediction probability of the i sample.

In order to improve the generalization ability of the model, L2 regularization and Dropout technology are introduced. In addition, in the choice of optimizer, Adam optimization algorithm is used, and its update rule is:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{7}$$

Where θ_t is the parameter vector, η is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ are the first and second moment estimates of the gradient, respectively.

4. Experiment and result analysis

4.1 Data set

The data set used in the experiment comes from several public databases, including TCGA(The Cancer Genome Atlas), String (Search Tool for the Retrieval of Interacting Genes/Proteins) and Kegg (Kyoto Encyclopedia of Genes and Genomes). After pretreatment, a biomolecular network with 2,500 nodes and 30,000 edges is finally obtained. Among them, nodes represent candidate molecules (such as genes and protein), and edges represent the interaction between them. It is not difficult to find from Table 1 that different types of biological data have different feature dimensions and sample distribution.

Table 1 Basic Statistics of the Dataset

Data Type	Number of Samples	Feature Dimension	Positive Sample Ratio	Negative Sample Ratio
Gene Expression Data	2,500	10,000	60%	40%
Protein Interaction Data	2,500	8,000	55%	45%
Metabolic Pathway Data	2,500	12,000	65%	35%

4.2 Experimental setup

In the study, the data set is randomly divided into training set, verification set and test set according to the proportion of 70%, 15% and 15%. Futhermore, the proposed algorithm is compared with several common methods. These commonly used methods include: random forest (RF), classical

classification algorithm support vector machine (SVM), standard GNN model GCN, and multi-task CNN(Multi-Task CNN). Table 2 shows the basic parameter settings of each method.

Table 2 Parameter Settings of Different Methods

Method Name	Parameter Settings
Random Forest	Number of Trees = 100, Max Depth = 10
SVM	Kernel Function = RBF, C = 1.0
GCN	Hidden Layer Dimension = 64, Activation Function = ReLU
Multi-Task CNN	Convolution Kernel Size = 3×3, Pooling Window = 2×2, Number of Fully Connected Layers = 2
Proposed Method	GNN + Multi-Task CNN, Node Sampling Strategy = Random Node Sampling & Random Walk Sampling

In this study, the following five evaluation indexes are selected, namely, accuracy, precision, recall rate, F1 score and AUC (Area Under Curve).

4.3 Model performance evaluation

Table 3 shows the performance evaluation results of each method on the test set. The proposed method is superior to other methods in all evaluation indexes, especially the accuracy and AUC values reach 89.7% and 92.6% respectively.

Table 3 Performance Evaluation Results on the Test Set

Method Name	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC Value (%)
Random Forest	78.3	76.2	75.8	76.0	81.2
SVM	80.1	79.4	78.9	79.1	82.5
GCN	84.5	83.7	83.2	83.4	87.8
Multi-Task CNN	86.2	85.4	85.0	85.2	89.3
Proposed Method	89.7	88.9	88.5	88.7	92.6

Table 4 further shows the influence of different node sampling strategies on the model performance. Combining the two sampling strategies can make full use of local and global information, thus further improving the model performance.

Table 4 Impact of Different Node Sampling Strategies on Model Performance

Sampling Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC Value (%)
Random Node Sampling	87.4	86.5	86.0	86.2	90.1
Random Walk Sampling	88.3	87.6	87.1	87.3	91.5
Combining Both Strategies	89.7	88.9	88.5	88.7	92.6

4.4 Result discussion

Table 5 lists some key molecules predicted by the model and their functional notes. These molecules play an important role in exosomes-mediated antioxidant process, and their prediction results are supported by the literature.

Table 5 Key Molecules Predicted by the Model and Their Functional Annotations

Molecule Name	Functional Description	Involved Pathway	Importance Score
SOD1	Superoxide dismutase, scavenges reactive oxygen species	Oxidative Stress Response	0.92

GPX1	Glutathione peroxidase, antioxidant defense	Nrf2/ARE Signaling Pathway	0.89
CAT	Catalase, decomposes hydrogen peroxide	Antioxidant Metabolism	0.87
NRF2	Transcription factor, activates antioxidant gene expression	Nrf2/ARE Signaling Pathway	0.85
HMOX1	Heme oxygenase, involved in iron metabolism and antioxidant	Redox Balance	0.83

The results show that the proposed algorithm can accurately identify the key molecules in the antioxidant pathway of exosomes and reveal the dynamic regulation relationship between them. For example, the synergistic effect of SOD1 and GPX1 may constitute the core link of exosomes' antioxidant mechanism, while Nrf2 further enhances the cell's defense ability by regulating the expression of downstream genes.

5. Conclusions

In this article, an exosome antioxidant pathway identification algorithm combining bioinformatics mining and GNN is proposed to analyze the exosome-mediated antioxidant mechanism. The research shows that it is difficult to fully explore the complex nonlinear relationship in biomolecular networks by traditional single method, but this study combines multi-group data with advanced GNN technology to overcome this limitation.

In this study, two sampling strategies, random node sampling and random walk sampling, are designed to capture local structural characteristics and long-range dependence respectively. The results show that the accuracy of the algorithm is 89.7% and the AUC value is as high as 92.6% on the test set, which is obviously superior to the traditional RF and SVM methods. In addition, the model successfully identified several key molecules such as SOD1, GPX1 and Nrf2, revealing their synergistic effect in the antioxidant process of exosomes. For example, SOD1 and GPX1 can directly relieve oxidative stress by scavenging reactive oxygen species, while Nrf2 can indirectly enhance the cellular defense ability by activating downstream gene expression.

Although some achievements have been made in this study, there are still some problems to be solved urgently. For example, the current data set is small, which may limit the generalization ability of the model; In addition, how to further optimize the calculation efficiency of GNN is also an important research direction. Future work will focus on expanding data sources, integrating more types of biological information such as single cell sequencing data, and exploring more complex GNN architectures such as graph attention network.

References

[1] Tan C Y, Ong H F, Lim C H, et al. Amogel: a multi-omics classification framework using associative graph neural networks with prior knowledge for biomarker identification[J]. BMC bioinformatics, 2025, 26(1): 1-27.

[2] Capecci E, Lobo J L, Laña I, et al. Modelling gene interaction networks from time-series gene expression data using evolving spiking neural networks[J]. Evolving Systems, 2020, 11(4): 599-613.

[3] Sun Y, Xie J, Yuan Z Y. Bioinformatics and Machine Learning Methods Identified MGST1 and QPCT as Novel Biomarkers for Severe Acute Pancreatitis[J]. Molecular biotechnology, 2024, 66(5):1246-1265.

[4] Yan A, Baricordi C, Nguyen Q, et al. IS-Seq: a bioinformatics pipeline for integration sites analysis with comprehensive abundance quantification methods[J]. BMC bioinformatics, 2023, 24(1): 286.

[5] Kolomeets M, Desnitsky V, Kotenko I, et al. Graph visualization: Alternative models inspired by bioinformatics[J]. Sensors, 2023, 23(7): 3747.

[6] Thareja P, Chhillar R S. A detailed survey on data mining based optimization schemes for bioinformatics applications[J]. ECS Transactions, 2022, 107(1): 4689.

[7] Dominic N, Elwirehardja G N, Pardamean B. Data Mining for the Global Multiplex Weekly Average Income Analysis[J]. Procedia Computer Science, 2023, 219: 52-59.

[8] Wang X, Wang Z, Zhang X, et al. Bioinformatics-assisted mining and design of novel pullulanase suitable for starch cold hydrolysis[J]. Journal of Biotechnology, 2025, 398: 106-116.

[9] Asadi F, Trinugroho J P, Hidayat A A, et al. Data mining for epidemiology: The correlation of typhoid fever occurrence and environmental factors[J]. Procedia Computer Science, 2023, 216: 284-292.

[10] Fei M, Lu C, Feng B, et al. Bioinformatics analyses and experimental validation of the role of phagocytosis in low-grade glioma[J]. Environmental Toxicology, 2024, 39(4): 2182-2196.

[11] Luo J, Wang J, Zhai H, et al. GCphase: an SNP phasing method using a graph partition and error correction algorithm[J]. BMC bioinformatics, 2024, 25(1): 267.

[12] Ma T, Wang H, Zhang L, et al. Graph classification based on structural features of significant nodes and spatial convolutional neural networks[J]. Neurocomputing, 2021, 423:639-650.

[13] Wang Y, Hou W, Sheng N, et al. Graph pooling in graph neural networks: Methods and their applications in omics studies[J]. Artificial Intelligence Review, 2024, 57(11): 294.