RESEARCH

Open Access

GNNMF: a multi-view graph neural network for ATAC-seq motif finding

Shuangquan Zhang[1], Xiaotian Wu[2], Zhichao Lian[1*], Chunman Zuo[3] and Yan Wang[2,4*]

Abstract

Background The Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) utilizes the Transposase Tn5 to probe open chromatic, which simultaneously reveals multiple transcription factor binding sites (TFBSs) compared to traditional technologies. Deep learning (DL) technology, including convolutional neural networks (CNNs), has successfully found motifs from ATAC-seq data. Due to the limitation of the width of convolutional kernels, the existing models only find motifs with fixed lengths. A Graph neural network (GNN) can work on non-Euclidean data, which has the potential to find ATAC-seq motifs with different lengths. However, the existing GNN models ignored the relationships among ATAC-seq sequences, and their parameter settings should be improved.

Results In this study, we proposed a novel GNN model named GNNMF to find ATAC-seq motifs via GNN and background coexisting probability. Our experiment has been conducted on 200 human datasets and 80 mouse datasets, demonstrated that GNNMF has improved the area of eight metrics radar scores of 4.92% and 6.81% respectively, and found more motifs than did the existing models.

Conclusions In this study, we developed a novel model named GNNMF for finding multiple ATAC-seq motifs. GNNMF built a multi-view heterogeneous graph by using ATAC-seq sequences, and utilized background coexisting probability and the iterloss to find different lengths of ATAC-seq motifs and optimize the parameter sets. Compared to existing models, GNNMF achieved the best performance on TFBS prediction and ATAC-seq motif finding, which demonstrates that our improvement is available for ATAC-seq motif finding.

Keywords Graph neural network, Coexisting probability, ATAC-seq motifs, Multi-view heterogeneous graph

*Correspondence:
Zhichao Lian
lzcts@163.com
Yan Wang
wy6868@jlu.edu.cn
[1] School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[3] Institute of Artificial Intelligence, Donghua University, Shanghai 201620, China
[4] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China

Background

Transcription factors (TFs) are proteins that can bind to DNA sequences and play a crucial role in gene-regulated networks [1], cell cycle regulation [2], and human diseases [3]. The region where TFs bind to DNA sequences is called transcription factor binding sites (TFBSs), which are short and conserved DNA fragments [4]. Transcription regulation is carried out by the interplay between TFs and TFBSs in DNA sequences, and identifying TFBSs aids us to reveal functions of TFs and the cause of human diseases [5]. The aligned TFBSs of the same TF can be defined as a regulatory DNA motif, which can be represented by a position weight matrix (PWM) [6]. Assays for Transposase-Accessible Chromatin using sequencing (ATAC-seq) can probe open chromatin via

Tn5, which is an effective method for locating TFBSs at a genome-wide level [7, 8]. Compared to Chromatin immunoprecipitation sequencing (ChIP-seq) and DNase I hypersensitive sites sequencing (DNase-seq) technologies, ATAC-seq can reveal more kinds of transcription factor binding regions [9]. An ATAC-seq footprint is a fragment of a DNA sequence that has not been cleaved due to the transcription factors binding. TOBIAS and Hint-ATAC are feasible tools for identifying ATAC-seq footprints, TOBIAS utilizes bias correction [10] and footprinting scores to locate footprints, and Hint-ATAC employs the hidden Markov model to locate footprints [11]. ATAC-seq footprints have been applied to TF network prediction [11], comparison of TF activity [12], identification of TFs enriched in peripheral blood mononuclear cells (PBMC)-specific peaks [13], and ATAC-seq motifs finding [14].

There are several special tools for ATAC-seq motif scanning, such as TRACE, chromVAR, SnapATAC [15–17]. They scan input sequences by using known motifs, which limits their ability to find ATAC-seq motifs. DL technology, including convolutional neural networks (CNNs) and the recurrent neural networks has achieved successes in protein-protein networks, gene-regulated networks, and motifs finding [18–20]. ATAC-seq motif finding includes two key steps: the first step is to predict an ATAC-seq sequence that contains TFBSs, *i.e.* TFBS prediction, and the second step is to find ATAC-seq motifs. Existing models such as scFAN, FactorNet, and DeepATAC employ CNNs to predict TFBSs and find motifs from ATAC-seq data [21, 22]. These models utilized convolutional kernels of CNNs to scan ATAC-seq sequences, and a fully connected layer to predict TFBSs. Due to the limitation of the width of convolutional kernels, these models only find motifs with fixed lengths. Moreover, these existing models do not consider the coexisting probability of TFBSs in an input sequence. In a related line of research, the graph neural network (GNN) can learn the key message from the graph-structured data [23]. GNNs can learn the embeddings of nodes via their neighboring nodes, and keep the connection of the graph unchanged and nodes' embedding can also enable graph-based explanation and reasoning. GNNs work on non-Euclidean structured data and have been shown to perform well on molecular structures [24], protein-protein networks [25], gene-gene networks [26], ATAC-seq motif finding [14], *etc.* MMGraph is an ATAC-seq motif predictor based on GNN and coexisting probability to find multiple motifs [14]. MMGraph utilizes ATAC-seq footprints to build a multi-view heterogeneous graph by defining coexisting, Hamming, jaccard, and inclusive edges. Compared with CNN-based models, MMGraph utilizes ATAC-seq footprints to build the multi-view heterogeneous graph and employs the coexisting probability to find multiple ATAC-seq motifs of different lengths. However, MMGraph still has certain defects. MMGraph sets the coexistence probability threshold to 0.5, which needs to be optimized. MMGraph uses only Hamming distance and coexisting probability to measure the relationships between k-mers, which ignored the correlation between k-mers.

To address the above issues, this study developed a novel model called GNNMF to find ATAC-seq motifs and predict TFBSs. GNNMF is based on MMGraph and has three improved aspects: built a multi-view heterogonous graph, which contains four kinds of edges (coexisting edges, similarity edges, jaccard edges, and inclusive edges) and two types of nodes (k-mer nodes and sequence nodes); defined the iterloss function to improve TFBS prediction accuracy; optimized the threshold of coexisting probability via defining coexisting probability between k-mer nodes in negative sequences. We tested the effectiveness of the GNNMF model on 200 human ATAC-seq datasets and 80 mouse ATAC-seq datasets across nine evaluation metrics, including the area of eight metrics radar (AEMR), precision, recall, F1_score, accuracy (ACC), specificity, the Matthews correlation coefficient (MCC), the area under the receiver operating characteristic curve (AUC), and the area under the precision-recall curve (PRC). According to the experimental results, GNNMF improved the ACC, MCC, and AUC by 9.6%, 18.2%, and 2.9%, respectively, on 200 human ATAC-seq datasets and by 2.36%, 6.35%, and 3.01%, respectively, on 80 mouse ATAC-seq datasets. In this study, AEMR is an overall score for evaluating the model's ability to predict TFBSs. GNNMF improved the AEMR by 4.92% and 6.81% on 200 human and 80 mouse ATAC-seq datasets, respectively. Moreover, among all the models, GNNMF find the most 385 and 662 significant motifs from human ATAC-seq data and mouse ATAC-seq data , respectively.

Methods

Data processing

The 280 ATAC-seq datasets, including 200 human ATAC-seq datasets and 80 mouse ATAC-seq datasets, were randomly downloaded from the ENCODE project (Supplementary Table S1). To obtain corrected footprints, footprints of each dataset are obtained via TOBIAS and Hint-ATAC tools with default parameters [11, 12]. All footprints are ranked in descending order by their scores. Then the top 1500 footprints are used to intersect with the footprints that Hint-ATAC obtains, and the intersected footprints of each dataset are applied to the following analysis.

Each intersected footprint is pruned with 101 bps around its center by the bedtools [27], which is set as a positive sequence. The TFBS prediction is a binary classification, so we shuffle all bases within a positive sequence as a negative sequence [28]. This paper gives a positive sequence a label of ‘1’, and gives a negative sequence a label of ‘0’. Thus, we can obtain an ATAC-seq sequence set S that includes n sequences.

For each dataset, 80% of S is set as training data, 10% of S is set as validation data, and the remaining S is set as testing data. Then, the sequence $s(\cdot)$ $(s(\cdot) \in S)$ is trimmed into k-mers $k(\cdot)$ by the step of one base to obtain a k-mer set $Ks(\cdot)(lenk = length(k(\cdot)))$. $s(\cdot)$ and k-mer $k(\cdot)$ will be utilized to build the multi-view heterogeneous graph G, where $s(\cdot)$ and $k(\cdot)$ are nodes.

Building the multi-view heterogeneous graph

The multi-view heterogeneous graph G contains two kinds of nodes (sequence node $s(\cdot)$ and k-mer node $k(\cdot)$), and four types of edges (coexisting edge, similarity edge, Jaccard edge and inclusive edge) (Fig. 1A). Coexisting edges (view 1) represent coexisting relationships between k-mer nodes in a sequence, and the weights of coexisting edges are calculated by the coexisting probability [14]. Similarity edges (view 2) represent relations between k-mer nodes and the weights of similarity edges are measured by Hamming distances [29]. Jaccard edges (view 3) represent relationships between k-mer nodes, and the weights of Jaccard edges are measured by the Jaccard correlation coefficient. Inclusive edges represent inclusive relations between a given sequence $s(\cdot)$ and its constituent k-mers, and weights of inclusive edges are measured by TF-IDF (term frequency-inverse document frequency) [30].

We let $Ks(i)$ be a k-mer set that contains all unique $k(\cdot)$ within $s(i)$. K is an m k-mer set that contains all unique $k(\cdot)$ in all $Ks(i)$ of S, where m is the total number of unique $k(\cdot)$ within S. Four edge types are defined to build G, where $s(\cdot)$ and $k(\cdot)$ are nodes.

Coexisting edges measure the coexisting probability between $k(\cdot)$ nodes. The coexisting edge weight between two k-mers $k(p)$ and $k(j)$ is calculated by Formula (1):

$$\begin{gathered} wco(p,j) = wco(j,p) = -\log(\frac{Q(k(p),k(j))}{P(k(p))P(k(j))}) \quad if \ nums(k(p),k(j)) > 0 \\ P(k(p)) = \frac{num(k(p))}{n}, \; Q(k(p),k(\mathrm{j})) = \frac{nums(k(p),k(j))}{n} \end{gathered} \tag{1}$$

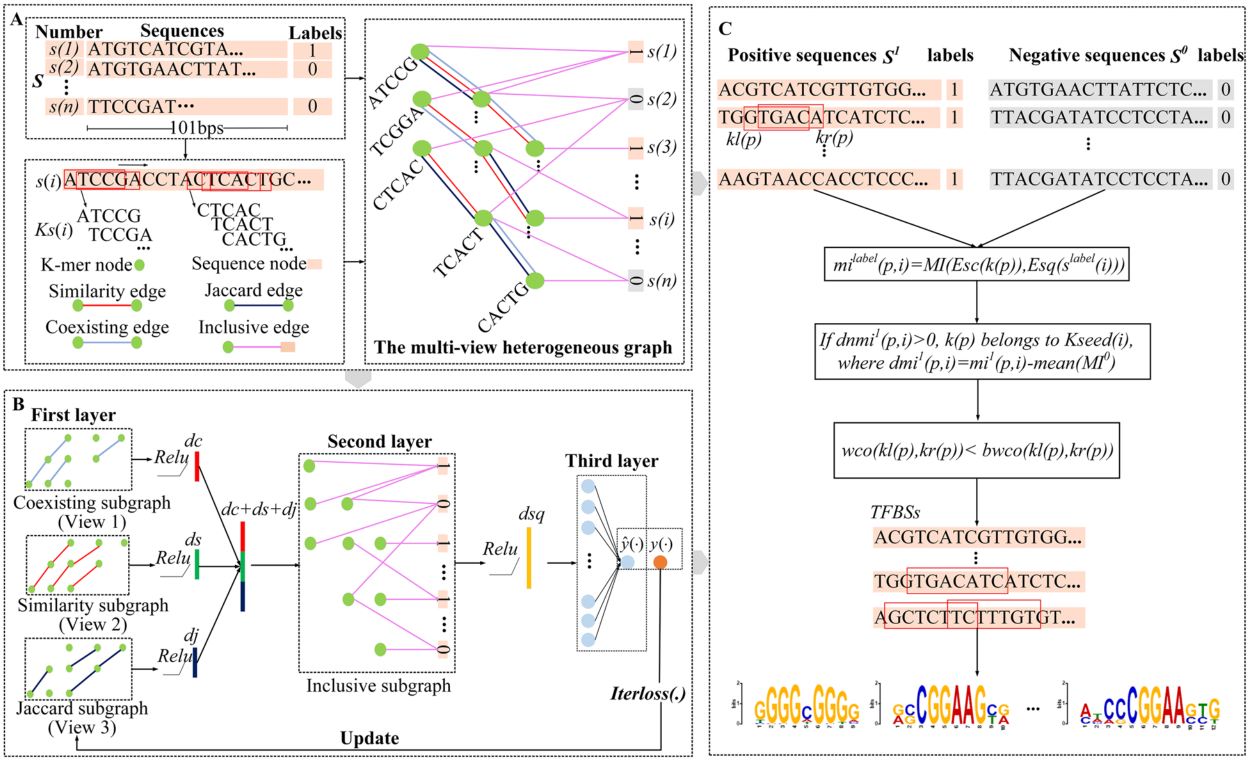

Fig. 1 The GNNMF framework

where Wco represents a $m \times m$ coexisting weight matrix, $num(k(\cdot))$ represents the total number of $Ks(\cdot)$ that contain $k(\cdot)$, and $nums(k(p), k(j))$ is the total number of $Ks(\cdot)$ that contain both $k(p)$ and $k(j)$.

Similarity edges measure mismatches between $k(\cdot)$ nodes using the Hamming distance [31]. The similarity edge weight between $k(p)$ and $k(j)$ is calculated by Formula (2):

$$wsim(p,j) = wsim(j,p) = Hamming(k(p), k(j)) \tag{2}$$

where $Wsim$ represents a $m \times m$ similarity weight matrix,$p, j \in [1, ..., m]$. $Ham\min g(\cdot,\cdot)$ represents the Hamming distance function.

Jaccard edges measure the correlation between $k(\cdot)$ nodes using the Jaccard correlation coefficient. The Jaccard edge weight between two nodes $k(p)$ and $k(j)$ is calculated by Formula (3):

$$wjac(p,j) = \frac{|Sk(k(p)) \cap Sk(k(j))|}{|Sk(k(p)) \cup Sk(k(j))|} \tag{3}$$

where $Sk(\cdot)$ presents a sequence set that contains all sequences, including $k(\cdot)$. $Wjac$ represents a $m \times m$ similarity weight matrix, $p, j \in [1, ..., m]$.

Inclusive edges measure the dependency degree between $s(\cdot)$ and $k(\cdot)$. The inclusive edge weight between $s(\cdot)$ and $k(\cdot)$ is calculated by transferring the concept of the term frequency-inverse document frequency (TF-IDF) [32] to Formula (4):

$$winclu(p,i) = tf(k(p), s(i)) \times \log(\frac{n}{num(k(p))}) \quad if\ k(p) \in Ks(i) \tag{4}$$

where $Winclu$ represents an $m \times n$ inclusive weight matrix,$tf(k(p), s(i))$ is the number of $k(p)$ existing in $s(i)$, $i \in [1, ..., n]$, and $p \in [1, ..., m]$.

GNNMF overview

In this study, we developed a three-layer GNN model named GNNMF to find ATAC-seq motifs. GNNMF decomposes the multi-view heterogeneous graph G into four subgraphs, *i.e.* a coexisting subgraph, similarity subgraph, Jaccard subgraph, and inclusive subgraph (Fig. 1B). The first layer of GNNMF is utilized to learn the embedding of $k(\cdot)$ as $Ek(k(\cdot)) \in \mathbb{R}^{dc+ds+dj}$ where dc, ds and dj are the embedding dimensions of $Esim(k(\cdot))(Esim(k(\cdot)) \in \mathbb{R}^{dc})$, $Eco(k(\cdot))$ $(Eco(k(\cdot)) \in \mathbb{R}^{ds})$ and $Ejac(k(\cdot))$ $(Ejac(k(\cdot)) \in \mathbb{R}^{dj})$ from the coexisting subgraph, similarity subgraph and Jaccard subgraph, respectively. The second layer learns the embedding of $s(\cdot)$ as $Es(s(\cdot))$ $(Es(s(\cdot)) \in \mathbb{R}^{dsq})$ from the inclusive subgraph, where dsq is the dimension of $Es(s(\cdot))$. The third layer is a fully connected layer, which is used to predict TFBSs.

We normalize Wco, $Wsim$ and $Wjac$ as the initial embedding of $k(\cdot)$ via Formula (5).

$$\begin{aligned} Ect(k(p)) &= (\sum_{j=1}^{m} Wco(p,j))^{-1} \times Wco(p,:) \\ Est(k(p)) &= (\sum_{j=1}^{m} Wsim(p,j))^{-1} \times Wsim(p,:) \\ Ejt(k(p)) &= (\sum_{j=1}^{m} Wjac(p,j))^{-1} \times Wjac(p,:) \end{aligned} \tag{5}$$

where $Ect(k(p))$ $(Ect(k(p)) \in \mathbb{R}^{dc})$ is the initial embedding of $k(p)$, based on Wco;$Est(k(p))(Est(k(p)) \in \mathbb{R}^{ds})$ is the initial embedding of $k(p)$, based on $Wsim$;$Ejt(k(p))$ $(Ejt(k(p)) \in \mathbb{R}^{dj})$ is the initial embedding of $k(p)$, based on $Wjac$.

GNNMF learns the embedding of $k(p)$ as $Eco(k(p))$ from the coexisting subgraph, and $Eco(k(p))$ is calculated by Formula (7).

$$eco(k(p)) = Ect(k(p)) \times Wco \times W^{co} \tag{6}$$

$$Eco(k(p)) = \mathrm{Re}LU(eco(k(p))) \tag{7}$$

where $W^{co}(W^{co} \in \mathbb{R}^{m \times dc})$ is a training weight matrix, which is randomly initialized. $\mathrm{Re}LU(\cdot)$ represents the rectified linear unit, which is an activation function.

GNNMF learns $Esim(k(p))$ from the similarity subgraph via Formula (9).

$$esim(k(p)) = Est(k(p)) \times Wsim \times W^{sim} \tag{8}$$

$$Esim(k(p)) = \mathrm{Re}LU(esim(k(p))) \tag{9}$$

where $Esim(k(p))$ is the embedding of $k(p)$ in the similarity subgraph. $W^{sim}(W^{sim} \in \mathbb{R}^{m \times ds})$ is a training weight matrix, which is randomly initialized.

GNNMF learns the embedding of $k(p)$ as $Ejac(k(p))$ from the Jaccard subgraph, and $Ejac(k(p))$ is calculated by Formula (11).

$$ejac(k(p)) = Ejt(k(p)) \times Wjac \times W^{jac} \tag{10}$$

$$Ejac(k(p)) = \mathrm{Re}LU(ejac(k(p))) \tag{11}$$

where $W^{jac}(W^{jac} \in \mathbb{R}^{m \times dj})$ is a training weight matrix that is randomly initialized.

Then, $Esim(k(\cdot))$, $Eco(k(\cdot))$ and $Ejac(k(\cdot))$ are concatenated as $Esc(k(\cdot))$ (Formula (12)), which is fed into the second layer.

$$Esc(k(p)) = concatenate([Eco(k(p)), Esim(k(p)), Ejac(k(p))]) \tag{12}$$

We stack $Esc(k(p))$ into a matrix Msc $(Msc \in \mathbb{R}^{m \times (ds+dc+dj)})$. GNNMF learns the embedding of $s(\cdot)$ as $Esq(s(\cdot))$ via the second layer, $Esq(s(\cdot))$ is calculated by Formula (14).

$$esq(s(i)) = Winclu(:, i)^T \times Msc \times W^{inclu} \tag{13}$$

$$Esq(s(i)) = \mathrm{Re}LU(esq(s(i))) \tag{14}$$

The third layer is a fully connected layer, which is used to predict TFBSs (Formula (16)).

$$\hat{y}(s(i)) = W(i, :) \times Esq(s(i)) + b(i) \tag{15}$$

$$\hat{y}(s(i)) = sigmoid(y'(s(i))) \tag{16}$$

where $W(W \in \mathbb{R}^{n \times dsq})$ is the training weight matrix, which is randomly initialized. b represents the bias, which is a $n \times 1$ vector, $sigmoid(\cdot)$ is a sigmoid function, and $\hat{y}(s(i))$ represents the predicted label of $s(i)$ by GNNMF.

The BCEloss is the binary cross-entropy loss, which is used to calculate the binary cross-entropy loss between the predicted probability and the true label. Based on BCEloss, we define the iterloss as the loss function of GNNMF [33]:

$$BCEloss(s(i)) = -[y(s(i)) \times \log \hat{y}(s(i)) + (1 - y(s(i))) \times \log(1 - \hat{y}(s(i)))] \tag{17}$$

$$Dloss(s(i))^{epoch-1} = BCEloss(s(i))^{epoch-1} \times sigmoid(BCEloss(s(i))^{epoch-1}) \tag{18}$$

$$\mathrm{iter}loss(i) = Dloss(s(i))^{epoch-1} + BCEloss(s(i))^{epoch} \tag{19}$$

where $y(s(i))$ represents the true label of sequence $s(i)$ (Formula (19)); $\hat{y}(s(i))$ represents the predicted label of sequence $s(i)$; and *epoch* represents the iteration, $BCEloss(s(i))^0 = 0$.

Find multiple motifs via coexisting probability between k-mers

GNNMF calculates the mutual information (MI) $mi(p, i)$ between $k(p)$ and $s(i)$ by using $Esc(k(\cdot))$ and $Esq(s(\cdot))$ (Fig. 1C). This study divided S into S^0 and S^1 by $label = 1/0$, as well as n^{label}, m^{label}, $s^{label}(\cdot)$, $Ks^{label}(\cdot)$ and K^{label}. GNNMF calculates the MI^{label} matrix by $mi(p, i)$, where $k(p) \in K^{label}$, $s(i) \in s^{label}(\cdot)$, $i \in [1, ..., n^{label}]$. We define $mean(MI^0)$ as the background noise of MI^1 and calculate the denoised MI matrix $dnMI^1$, where $mean(MI^0) = (\sum_{p=1}^{m^0} \sum_{i=1}^{n^0} mi^0(p, i))/(m^0 \times n^0)$ and $dnmi^1(p, i) = mi^1(p, i) - mean(MI^0)$. We define the k-mer seed set $Kseed(i)$ that contains all unique $k(p)$ in $s^1(i)$, when $dnmi^1(p, i) > 0, i \in [1, ..., n^1]$. Then, we locate the interval of $k(p)$ in $s^1(i)$ as $itk(p) = [strk(p), strk(p) + lenk - 1]$, where $k(p) \in Kseed(i), strk(p)$ is the starting position of $k(p)$ in $s^1(i)$. If multiple $k(p)$ exist in $s^1(i)$, there will be multiple $itk(p)$. We obtain two $k(\cdot)$ centered on $itk(p)$ as $kl(p) = [ck(p) - lenk + 1, ck(p)]$ and $kr(p) = [ck(p){+}1, ck(p) + lenk]$, where $ck(p) = strk(p) + \lceil (lenk - 1)/2 \rceil$ and $kl(p), kr(p) \in K$.

We calculate the coexisting probability $wco(kl(p), kr(j))$ and background probability $bwco(kl(p), kr(p))$ between $kl(p)$ and $kr(p)$, respectively. If $wco(kl(p), kr(p)) < bwco(kl(p), kr(p))$, which means that $kl(p)$ and $kr(p)$ are strongly related, GNNMF merges $kl(p)$ and $kr(p)$ to a TFBS as $tfbs(p) = [ck(p) - lenk{+}1, ck(p) + lenk]$. For all $k(\cdot) \in Kseed(i)$, we can find multiple candidates $tfbs(i)$ in $s^1(i)$. If multiple $tfbs(i)$ overlap, they are merged into a longer TFBS in $s^1(i)$. Finally, GNNMF can find multiple $tfbs(\cdot)$ with different lengths.

Evaluation metrics

Our evaluation metrics include precision, recall, F1_score, ACC, specificity, MCC, AUC, PRC, and AEMR [34]. AEMR is the area of a radar chart, which can be generated that consists of eight equiangular spokes with each spoke representing one of the scores defined above [14]. The higher the AEMR score is, the better the performance of model for TFBS prediction.

All the evaluation metrics are calculated via the following formulas:

$$Precision = \frac{TP}{TP + FN} \tag{20}$$

$$\mathrm{Recall} = \frac{TP}{TP + TN} \tag{21}$$

$$\frac{2}{f1\_score} = \frac{1}{precision} + \frac{1}{recall} \tag{22}$$

$$Specificity = \frac{TN}{FP + TN} \tag{23}$$

$$ACC = \frac{TN + TP}{FP + FN + TP + TN} \tag{24}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (FP + FN) \times (TN + FP) \times (TN + FN)}} \tag{25}$$

where TP, TN, FP and FN represent the number of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN), respectively.

The AUC is the area under the receiver operating characteristic curve, and the PRC is the area under the precision-recall curve, which is a value between 0 and 1.

$$O_{i,i+1} = \frac{1}{2} R_i \times R_{i+1} \times \sin(\frac{\pi}{4}) \quad i = 1, ..., 8 \tag{26}$$

$$\begin{aligned} &R = [precision, recall, F1\_score, ACC, specificity, MCC, AUC, PRC, precision] \\ &AEMR = sum(O_{i,i+1}) \quad i = 1, ..., 8 \end{aligned} \tag{27}$$

Experiment settings

This study trained GNNMF for 30 epochs using the Adam optimizer [35] (Fig. 2). The hyperparameters of GNNMF included the learning rate, *dsim*, *dco*, *djac*, *dsq* and *lenk*, and their ranges were tested in this study (Supplementary Table S2). We utilized the grid search to choose the optimal parameters and applied the combinations of those hyperparameters to GNNMF model. The AUC is used to measure the performance of GNNMF with different combinations on the testing set. The learning rate of GNNMF was set as 0.001 and was adjusted by the natural exponential decay with 0.001. When we trained GNNMF on the training set, the sequence nodes in validation set and the testing set were masked. The *dsim*, *dco* and *djac* were set as 50, and *dsq* was set as 150 in our experiment. Considering the computational complexity and the performance of GNNMF, we set *lenk* = 5. The scFAN, DeepATAC, FactorNet, MMGraph, and MMGraph+jac models are used as comparison tools, where MMGraph+jac combines MMGraph and jaccard graph. The evaluation metrics, including precision, recall, F1_score, ACC, specificity, MCC, AUC, PRC, and AEMR, are utilized to evaluate models' performance on TFBS prediction. To explain the GNNMF model, ATAC-seq motifs are used to show features that GNNMF learned. We matched the found motifs to motifs that the HOCOMOCO database contains via the TOMTOM tool.

Results

TFBS prediction

In this study, we utilized nine metrics to evaluate the performance of all the models on 200 human ATAC-seq datasets and 80 mouse ATAC-seq datasets (Fig. 3). FactorNeT, scFAN, DeepATAC, MMGraph, and MMGraph+jac were selected as comparison tools, because they achieved feasible performances in the previous study. In the light of results, AEMRs of models based on the GNN model are higher than those of the models based on the CNN model, which demonstrates that the

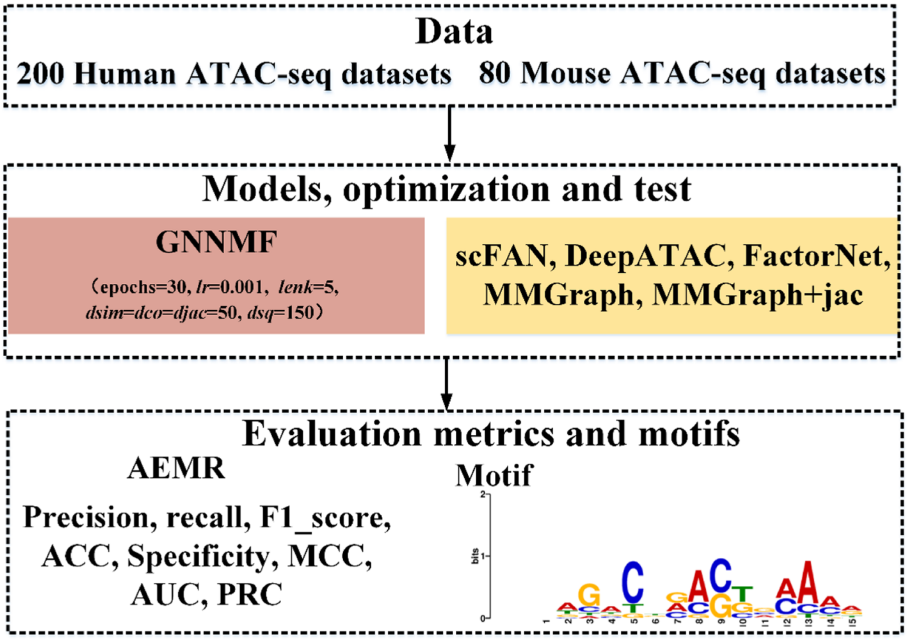
Data
200 Human ATAC-seq datasets
80 Mouse ATAC-seq datasets
Models, optimization and test
GNNMF
(epochs=30, lr=0.001, lenk=5,
dsim=dco=djac=50, dsq=150)
scFAN, DeepATAC, FactorNet,
MMGraph, MMGraph+jac
Evaluation metrics and motifs
AEMR
Precision, recall, F1_score,
ACC, Specificity, MCC,
AUC, PRC
Motif
bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Fig. 2 The workflow of the experimental setting

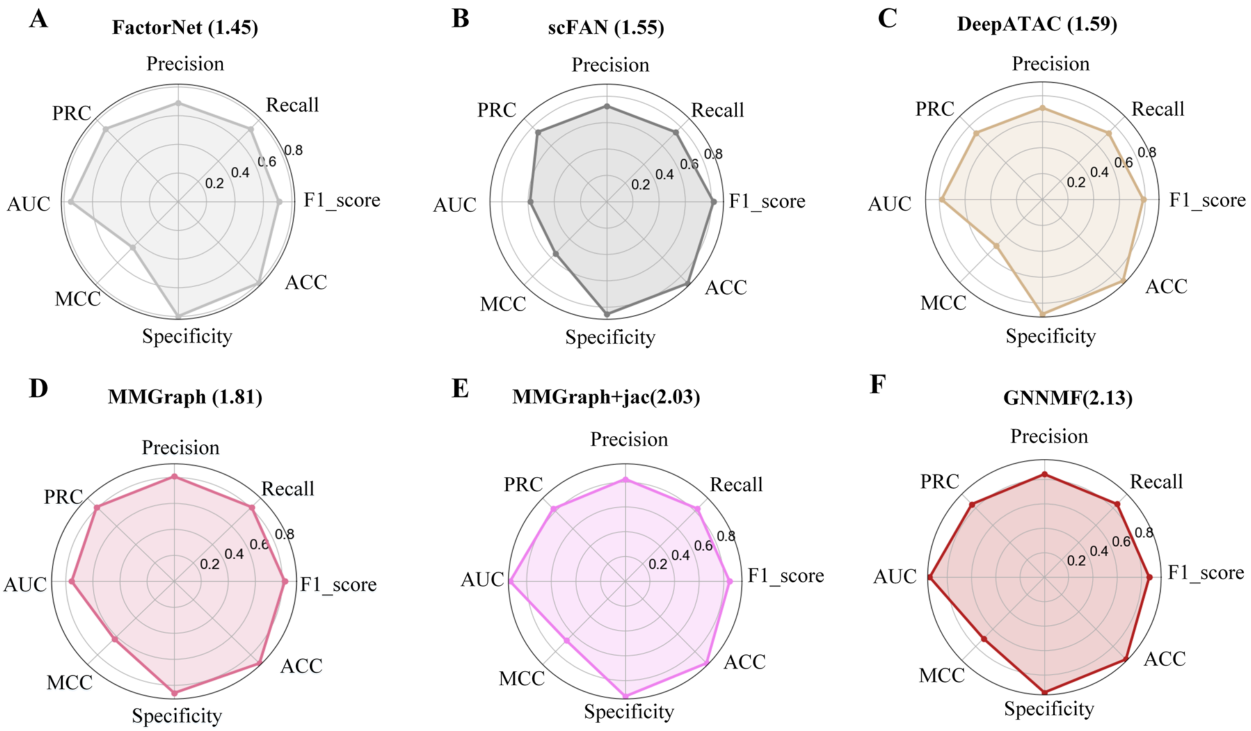

Fig. 3 AEMRs of six models on 200 human ATAC-seq datasets

GNN model is more suitable for ATAC-seq TFBS prediction. On 200 human ATAC-seq datasets, GNNMF achieved the highest average AEMR score (2.13). Compared to existing models, GNNMF improved the AEMR by 4.92%. Among models based on CNNs, DeepATAC obtained an AEMR of 1.59 , which is higher than those of scFAN and FactorNeT. On 80 mouse ATAC-seq datasets, GNNMF achieved consistent results on 200 human ATAC-seq datasets, and improved the AEMR by 6.81% (Supplementary Figure S1). On 280 ATAC-seq datasets, GNNMF yielded the highest AEMR score, which indicated that our improvement was efficient.

On 200 human ATAC-seq datasets, GNNMF obtained the highest scores for all eight metrics and achieved average precision, recall, F1_score, ACC, specificity (Supplementary Figure S2), MCC, AUC, and PRC values of 0.863, 0.846, 0.843, 0.845, 0.944, 0.709, 0.942, and 0.949, respectively. In particular, GNNMF improved the ACC and MCC by 2.13% and 5.08%, respectively. Moreover, on 80 mouse ATAC-seq datasets, GNNMF achieved average precision, recall, F1_score, ACC, specificity, MCC, AUC, and PRC values of 0.844, 0.831, 0.828, 0.830, 0.917, 0.676, 0.931, and 0.946, respectively (Supplementary Table S3). Compared to existing models, GNNMF improved the ACC and MCC by 2.36% and 6.35%, respectively.

Finding multiple ATAC-seq motifs

ATAC-seq can reveal all opening chromatin, which means that there are multiple motifs in an ATAC-seq dataset. In this study, we developed the GNNMF model to find ATAC-seq motifs by employing the GNN and coexisting probability. Existing DL models to find ATAC-seq motifs, including FactorNet, scFAN, DeepATAC, and MMGraph, were selecteds as comparison tools. We tested the above models on 280 ATAC-seq datasets including 200 human ATAC-seq datasets and 80 mouse ATAC-seq datasets. This study matches the found motifs to the HOCOMOCO database via the TOMTOM v5.1.0 tool. The *p*-value less than 0.05 indicated that the found motifs are significant. All motifs that each model found are listed in Table 1, 401 human ATAC-seq motifs were found. Among all the models, GNNMF found the most ATAC-seq motifs. Moreover, we tested all the models on 80 mouse ATAC-seq datasets, and the number of motifs in which each model found is listed in Supplementary Table S4. The *p*-value depicts the degree to which the found motifs are significant. The *p*-values of the found motifs of each model are shown in the violin plot, and the median value shows the model's ability to find ATAC-seq motifs (Fig. 4). GNNMF achieved the highest median value among all the models. Based on motif finding results, GNNMF found the most motifs among all the models. The GNN-based models outperformed the CNN-based models, which indicated that the

Table 1 Count of motifs that each model found on 200 human ATAC-seq datasets

Models	FactorNet	scFAN	DeepATAC	MMGraph	MMGraph+jac	GNNMF
Motifs.no	270	291	281	375	375	385

GNN is a feasible algorithm for finding ATAC-seq motifs. GNNMF is based on MMGraph, which is improved via the Jaccard edge, background probability, and the iterloss function. Our results demonstrated that our improvements were conducive to finding ATAC-seq motifs.

Discussion

This study developed a GNN model called GNNMF by improving the MMGraph model. We improved the MMGraph by using the Jaccard edge, iterloss function, and background probability. GNNMF decomposed the multi-view heterogeneous graph into four subgraphs, and employed a three-layer GNN model to predict TFBSs. The first layer of GNNMF was used to learn the embedding of k-mer nodes via the similarity subgraph, Jaccard subgraph, and coexisting subgraph. The second layer was used for the embedding of sequence nodes via the inclusive subgraph, and the third layer was used to predict TFBSs. GNNMF utilizes embeddings of k-mer nodes and sequence nodes to define k-mer seeds and employs the coexisting probability to find multiple motifs with different lengths. Existing models to find ATAC-seq motifs, including FactorNet, scFAN, DeepATAC, MMGraph and MMGraph +jac, were selected as comparison tools. We tested all the models on 200 human ATAC-seq datasets and 80 mouse ATAC-seq datasets. We utilized the precision, recall, F1_score, specificity, ACC, MCC, AUC, and PRC to evaluate the models' ability in TFBS prediction. In light of our results, GNNMF achieved the highest precision, recall, F1_score, ACC, specificity, MCC, AUC, and PRC of 0.863, 0.846, 0.843, 0.845, 0.944, 0.709, 0.942, and 0.949, respectively, on 200 human datasets. On 88 mouse datasets, our proposed model obtained the highest precision, recall, F1_score, ACC, specificity, MCC, AUC, and PRC values of 0.844, 0.831, 0.828, 0.830, 0.917, 0.676, 0.931, and 0.946, respectively. Moreover, the GNN-based model achieved higher AEMR scores than models based on CNN. Our results demonstrated that the GNN has more potential for TFBS prediction on ATAC-seq datasets. Compared to MMGraph+jac, GNNMF utilizes the iterloss to predict TFBSs. The AEMR of GNNMF was higher than that of MMGraph+jac, which indicated the iterloss has an advantage over BCEloss in TFBS prediction.

GNNMF utilized the embedding of k-mer nodes and sequence nodes, and calculated coexisting probability between k-mer nodes to find multiple motifs with different lengths. MMGraph and MMGraph+jac used the same way to set the background probability threshold, and they set the background probability threshold to 0.5. However, GNNMF defines the background probability threshold via negative sequences. Our results indicated that GNNMF found more and higher quality motifs than MMGraph and MMGraph+jac. Therefore, the background probability threshold of the negative sequences

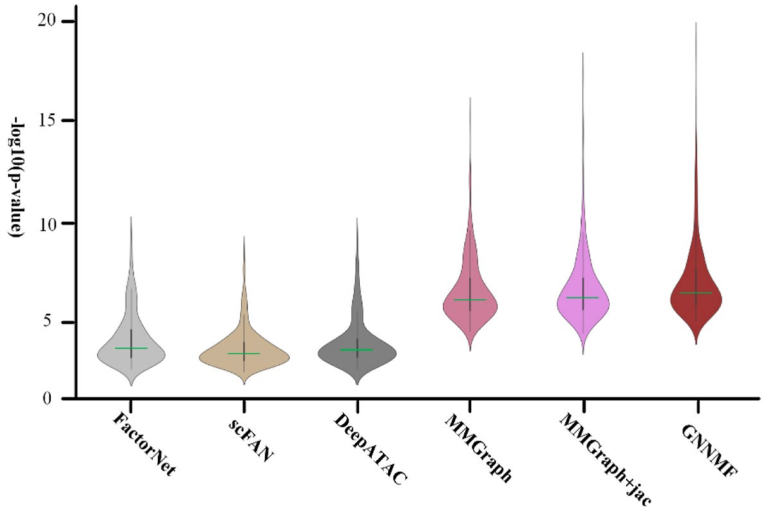

Fig. 4 *p*-values of the found motifs of six models on 200 human ATAC-seq datasets

aided GNNMF in finding ATAC-seq motifs. FactorNet, scFAN, and DeepATAC are all based on CNNs, they found ATAC-seq motifs by convolutional kernels in the first layer. But these models only found the fixed length of motifs, and the found motifs are similar. However, compared with models based on CNNs, models based on GNNs have great advantages. We utilized all the models to find ATAC-seq motifs and used p-value to evaluate the motifs' quality. The results demonstrated that GNNMF is the best model for finding multiple ATAC-seq motifs.

GNNMF employs a three-layer GNN to predict whether given sequences are bound by TFs. Some novel algorithms can be applied in TFBS prediction, such as graph attention networks [36], graphGAN [37], and graph autoencoders [38]. In the heterogeneous graph, there are many relationships between nodes and each kind of relationship represents the importance between two nodes. The attention mechanism can allocate and update the different weights to the nodes and edges of the heterogeneous graph, during the training process of the model. Moreover, TFs bind indirectly to motifs of other TFs, which co-regulate targeted gene expression [39]. The cooperation of TFs acts as a vital role in the process of human disease [40]. ATAC-seq data can detect open-accessible DNA regions by probing open chromatin, meaning that ATAC-seq data contain multiple TFs [41]. By analyzing ATAC-seq data, we revealed interactions between TFs, and explored the inducement of human disease. Therefore, GNNMF is a potential tool for studying the cooperation among different TFs.

Conclusions

In this study, we developed a novel model named GNNMF for finding multiple ATAC-seq motifs. GNNMF built the multi-view heterogeneous graph by using ATAC-seq sequences, employed a three-layer of GNN to predict TFBSs, and utilized coexisting probability to find ATAC-seq motifs. We conducted experiments on 200 human and 80 mouse ATAC-seq datasets to analyze the effectiveness of the proposed method. Our evaluation metrics included precision, recall, F1_score, ACC, specificity, MCC, AUC, PRC, and AEMR. In particular, GNNMF improved the AEMR by 4.92% and 6.81% on 200 human and 80 mouse ATAC-seq datasets, respectively. Meanwhile, GNNMF found multiple motifs from ATAC-seq data through the coexisting probability between k-mers. Regarding ATAC-seq data, our proposed method found more and higher quality motifs, which demonstrated methods based on coexisting probability of k-mers are more efficient than DL models. As a result, GNNMF achieved better performance than a few state-of-the-art methods. This study made great contributions to finding motifs from ATAC-seq data.

Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
TFBSs	transcription factor binding sites
DL	Deep learning
CNNs	convolutional neural networks
GNN	Graph neural network
TFs	Transcription factors
PWM	position weight matrix
ChIP-seq	Chromatin immunoprecipitation sequencing
DNase-seq	DNase I hypersensitive sites sequencing
AEMR	area of eight metrics radar
ACC	accuracy
MCC	Matthews correlation coefficient
AUC	Area under the receiver operating characteristic curve
TF-IDF	Term frequency-inverse document frequency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10218-0.

Supplementary Material 1.

Acknowledgements

We gratefully acknowledge generous support for this research from our team.

Authors' contributions

SZ conceived the idea. SZ and XW processed the data and conducted simulation and real dataset experiments. CZ and SZ wrote the manuscript. ZL and YW revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China [62302218, 32300523], the National Natural Science Foundation of China [62072212].

Availability of data and materials

The datasets are provided within Supplementary Table S1, which can be downloaded from the ENCODE project. GNNMF is available at https://github.com/zhangsq06/GNNMF.git.

Declarations

Ethics approval and consent to participate

Humans, animals or plants have not been directly used in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 29 November 2023 Accepted: 12 March 2024
Published online: 21 March 2024

References

1. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res. 2003;31(4):1234–44.
2. Joaquin Á, Watson R. Cell cycle regulation by the B-Myb transcription factor. Cell Mol Life Sci. 2003;60:2389–401.
3. Vishnoi K, Viswakarma N, Rana A, Rana B. Transcription factors in cancer development and therapy. Cancers (Basel). 2020;12(8):1–32.

4. Wang Y, Zhang S, Ma A, Wang C, Ma Q. Assessing deep learning algorithms in cis -regulatory motif finding based on genomic sequencing data. Briefings in Bioinformatics. 2020;23(1):1–10.
5. Brenner S, Wersinger C, Gasser T. Transcriptional regulation of the α-synuclein gene in human brain tissue. Neurosci Lett. 2015;599:140–5.
6. Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics. 2006;22(14):e454–63.
7. Bajic M, Maher KA, Deal RB. Identification of open chromatin regions in plant genomes using ATAC-Seq. Methods Mol Biol. 2018;1675:183–201.
8. Sun Y, Miao N, Sun T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. Hereditas. 2019;156(1):1–9.
9. Ma S, Zhang Y. Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. Mol Biomed. 2020;1:1–13.
10. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun. 2020;11(1):4267.
11. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20:1–21.
12. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun. 2020;11(1):1–11.
13. Youn A, Marquez EJ, Lawlor N, Stitzel ML, Ucar D. BiFET: sequencing Bias-free transcription factor Footprint Enrichment Test. Nucleic Acids Res. 2018;47(2):e11–e11.
14. Zhang S, Yang L, Wu X, Sheng N, Fu Y, Ma A, Wang Y. MMGraph: a multiple motif predictor based on graph neural network and coexisting probability for ATAC-seq data. Bioinformatics. 2022;38(19):4636–8.
15. Ouyang N, Boyle AP. TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence. Genome Res. 2020;30(7):1040–6.
16. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14(10):975–8.
17. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun. 2021;12(1):1337.
18. Lei H, Quan C. A Shortest Dependency Path Based Convolutional Neural Network for Protein-Protein Relation Extraction. BioMed Res Int, 2016, (2016-7-14). 2016, 2016:1-9.
19. Wang Y, Zhang S, Yang L, Yang S, Ma Q. Measurement of conditional relatedness between genes using fully convolutional neural network. Front Genet. 2019;10:1009.
20. Yang J, Ma A, Hoppe AD, Ang CW, Li Y, Zhang C, Wang Y, Liu B, Ma Q. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res. 2019;15:15.
21. Fu L, Zhang L, Dollinger E, Peng Q, Nie Q, Xie X. Predicting transcription factor binding in single cells through deep learning. Science Advances. 2020;6(51):eaba9031.
22. Quang D, Xie XS. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–47.
23. Scarselli F, Yong SL, Gori M, Hagenbuchner M, Tsoi AC, Maggini M. Graph Neural Networks for Ranking Web Pages. 2005.
24. Zhang S, Liu Y, Xie L: Molecular Mechanics-Driven Graph Neural Network with Multiplex Graph for Molecular Structures. 2020.
25. Zhang D, Kabuka M. Multimodal deep representation learning for protein interaction identification and protein family classification. BMC Bioinformatics. 2019;20(16):1–14.
26. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks. Sci Rep. 2022;12(1):1–12.
27. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
28. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
29. Grabowski S, Kowalski TM: Algorithms for all-pairs Hamming distance based similarity. Software: Practice and Experience 2021.
30. Mu'tasem J, Salim N. Stock market prediction based on term frequency-inverse document frequency. J Econ Bus Mgmt. 2016;4(3):183–7.
31. Norouzi M, Fleet DJ, Salakhutdinov RR. Hamming distance metric learning. Adv Neural Inform Process Syst. 2012;25:1–9.
32. Yun-tao Z, Ling G, Yong-cheng W. An improved TF-IDF approach for text classification. J Zhejiang Univ-Sci A. 2005;6(1):49–55.
33. Malhotra R, Shakya A, Ranjan R, Banshi R. Software defect prediction using binary particle swarm optimization with binary cross entropy as the fitness function. In: Journal of Physics: Conference Series: 2021: IOP Publishing; 2021: 012003.
34. Mancini A, Vito L, Marcelli E, Piangerelli M, De Leone R, Pucciarelli S, Merelli E. Machine learning models predicting multidrug resistant urinary tract infections using "DsaaS." BMC Bioinformatics. 2020;21(10):1–12.
35. Mehta S, Paunwala C, Vaidya B. CNN based traffic sign classification using Adam optimizer. 2019 international conference on intelligent Computing and Control Systems (ICCS). 2019:1293–8.
36. Velikovi P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. 2017.
37. Wang H, Jia W, Wang J, Miao Z, Guo M: GraphGAN: Graph Representation Learning with Generative Adversarial Nets. IEEE Transactions on Knowledge and Data Engineering. 2017, PP(99).
38. Hong C, Chen L, Liang Y, Zeng Z. Stacked Capsule Graph Autoencoders for geometry-aware 3D head pose estimation. Comp Vision Image Understand. 2021;1:103224.
39. Nie Y, Shu C, Sun X. Cooperative binding of transcription factors in the human genome. Genomics. 2020;112(5):3427–34.
40. Di Malta C, Cinque L, Settembre C. Transcriptional regulation of autophagy: Mechanisms and diseases. Front Cell Dev Biol. 2019;7(114).
41. Doganli C, Sandoval M, Thomas S, Hart D. Assay for Transposase-Accessible Chromatin with High-Throughput Sequencing (ATAC-Seq) Protocol for Zebrafish Embryos. Methods Mol Biol. 2017;1507:59.

Publisher's Note