Genome analysis

**MMGraph: a multiple motif predictor based on graph neural network and coexisting probability for ATAC-seq data**

Shuangquan Zhang, Lili Yang, Xiaotian Wu, Nan Sheng, Anjun Ma and Yan Wang

1Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China, 2Department of Obstetrics, The First Hospital of Jilin University, Changchun 130012, China, 3School of Artificial Intelligence, Jilin University, Changchun 130012, China, 4Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, UK and 5Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 19, 2022; revised on July 23, 2022; editorial decision on August 13, 2022; accepted on August 21, 2022

Abstract

**Motivation:** Transcription factor binding sites (TFBSs) prediction is a crucial step in revealing functions of transcription factors from high-throughput sequencing data. Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) provides insight on TFBSs and nucleosome positioning by probing open chromatic, which can simultaneously reveal multiple TFBSs compare to traditional technologies. The existing tools based on convolutional neural network (CNN) only find the fixed length of TFBSs from ATAC-seq data. Graph neural network (GNN) can be considered as the extension of CNN, which has great potential in finding multiple TFBSs with different lengths from ATAC-seq data.

**Results:** We develop a motif predictor called MMGraph based on three-layer GNN and coexisting probability of k-mers for finding multiple motifs from ATAC-seq data. The results of the experiment which has been conducted on 88 ATAC-seq datasets indicate that MMGraph has achieved the best performance on area of eight metrics radar score of 2.31 and could find 207 higher-quality multiple motifs than other existing tools.

**Availability and implementation:** MMGraph is wrapped in Python package, which is available at https://github.com/zhangsq06/MMGraph.git

**Contact:** wy6688@jlu.edu.cn

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1. Introduction

Identifying the corresponding transcription factor binding sites (TFBSs) which are short and conserved DNA motifs, is crucial for studying transcription factor (TF) regulations. The footprints of Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) can simultaneously reveal more kinds of TFBSs than traditional sequencing technologies (Bentsen et al., 2020). Furthermore, a graph neural network (GNN) can be considered as the extension of convolutional neural network, which learns the representation of target nodes via their neighbors (Feng et al., 2020). GNN has great potential in finding multiple motifs from ATAC-seq data. In this study, we develop a robust and effective tool, named MMGraph, for finding multiple motifs from ATAC-seq data.

2. Materials and methods

MMGraph is based on GNN and coexisting probability of k-mers, where the coexisting probability represents the degree of association between k-mers. It consists of three components (Fig. 1): (i) a heterogeneous graph; (ii) a three-layer GNN model to get embeddings of k-mers and sequences; and (iii) coexisting probability calculation for finding multiple motifs.

2.1 Constructing the heterogeneous graph

The sequence set S contains n sequences s(·), s(·) ∈ S. If s(i) contains TFBSs, we set label of s(i) as positive ‘1’, otherwise as negative ‘0’, i ∈ [1, . . . , n]. s(·) is trimmed into the same length fragments of
Fig. 1. The whole MMGraph workflow consists of three steps

2. Building the GNN model

MMGraph decomposes the heterogeneous graph into three subgraphs, i.e., similarity graph, coexisting graph and inclusive graph. Then, it uses the labels of \( s(\cdot) \) as targets to train a three-layer of GNN model. The first layer learns the concatenated embedding of \( k(\cdot) \) as \( Esc(k(\cdot)) \in \mathbb{R}^{d_{k}} \), where \( ds \) and \( dc \) are the embedding dimensions of \( Es(k(\cdot)) \in \mathbb{R}^{ds} \) and \( Ee(k(\cdot)) \in \mathbb{R}^{dc} \) from similarity graph and coexisting graph respectively. The second layer learns the embedding of \( s(\cdot) \) as \( Esq(s(\cdot)) \in \mathbb{R}^{dsq} \) from inclusive graph, where \( dsq \) is the dimension of \( Esq(s(\cdot)) \). The third layer is a fully connected layer, which identifies whether a label of \( s(\cdot) \) is ‘1’ that contains TFBSs or not by threshold 0.5 (Supplementary material).

3. Finding multiple motifs

After getting the embeddings of \( k(\cdot) \) and \( s(\cdot) \), we calculate mutual information (MI) of \( k(\cdot) \) and \( s(\cdot) \) by their corresponding embedding \( Esc(k(\cdot)) \) and \( Eq(s(\cdot)) \). The \( mi(p,i) \) measures the information that \( k(p) \) and \( s(i) \) shared. In order to indicate MI for different \( s(\cdot) \) with labels, we subdivide \( S \) into \( S^{label} \) by sequences’ label of ‘1’ and ‘0’, as well as \( nl^{label} \). Thus, \( K_{s}^{label}(\cdot) \) and \( K^{label} \) for \( n, m, s(\cdot), K_{s}(\cdot) \) and \( K \). The whole procedure of motifs finding includes four steps:

1. Obtain the \( M^{label} \) matrix by calculating \( mi(p,i) \), where \( k(p) \in K_{s}^{label} \), \( s(i) \in s^{label}(\cdot), i \in [1, \ldots, n^{label}], label = 1/0 \).
2. Get the denoised \( dnM^{i} \) by \( dnmi(p,i) = mi(p,i) - mean(M^{i}) \), where the background noise \( mean(M^{i}) = \sum Q_{n}^{i} \sum_{p} mi(q,p)/(m^{i} \times n^{i}) \).
3. get the \( K_{seed}(\cdot) \) that contains all unique \( k(p) \) in \( s^{i}(\cdot) \), s.t. \( dmni(p,i) > 0, i \in [1, \ldots, n^{i}] \), then locate the interval of \( k(p) \) in \( s^{i}(\cdot) \) as \( itvk(p) = [startk(p) + lenk(p) - 1], where k(p) \in K_{seed}(\cdot), startk(p) \) is the starting position of \( k(p) \) in \( s^{i}(\cdot) \). If multiple \( k(p) \) exist in \( s^{i}(\cdot) \), there will be multiple \( itvk(p) \).
4. then get two \( k(p) \) centered on \( itvk(p) \) as \( k(p) = [ck(p) + lenk(p) - 1, ck(p)] \), \( lenk(p) = lenk(p) + 1, ck(p) + lenk(p) \), \( lenk(p) = [ck(p) + 1, ck(p) + lenk(p)] \), \( tk(p) = [ck(p) + 1, ck(p) + lenk(p)] \), \( k(p) \in K \). If \( wco(k(p), k(v)) > -log(0.5), which means that k(p) \) and \( k(v) \) have a strong relation, we merge \( k(p) \) and \( k(v) \) to a candidate TFBS as \( tfbs(p) = [ck(p) + lenk(p) + 1, ck(p) + lenk(p)], otherwise, k(p) \) and \( k(v) \) will not be considered. For all \( k(\cdot) \in K_{seed}(\cdot) \), we can find multiple candidates \( tfbs(\cdot) \) in \( s^{i}(\cdot) \). If multiple \( tfbs(\cdot) \) overlapped, they may be merged to a longer TFBS in \( s^{i}(\cdot) \). Finally, we can find multiple \( tfbs(\cdot) \) with different lengths.

3. Experiments and results

The experiment has been conducted on 88 ATAC-seq datasets. For each sequence set, sequences in this set were acquired by ATAC-seq footprints, which were all positive sequences with label ‘1’. We randomly shuffled all bases within a positive sequence to generate a negative sequence with label ‘0’ (Alipanahi et al., 2015). The negative sequence did not contain TFBSs, but with the same GC content as the positive sequence. The area of eight metrics radar (AEMR) was used to evaluate tools’ capability in identifying whether the sequence contains TFBSs (Zhang et al., 2022), and the \( P \)-value was used to measure the quality of tools’ found motifs. Our results show that MMGraph achieved the highest AEMR score of 2.31 while \( lenk = 5 \), and could find 207 higher-quality motifs compare to four comparison tools (Supplementary data).

Funding

This work was supported by the National Natural Science Foundation of China [62072212]; Jilin province project [20220508125RC] and the Chinese Postdoctoral Science Foundation [2021M691211].

Conflict of Interest: The authors declare that they have no conflict of interest.
Data availability
The data underlying this article are available in its online supplementary material.

References
Alipanahi, B. et al. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol., 33, 831–838.
Bentsen, M. et al. (2020) ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat. Commun., 11, 1–11.

Colonnese, S. et al. (2021) Protein-Protein Interaction Prediction via Graph Signal Processing. In: IEEE Access, vol. 9, pp. 142681–142692. https://doi.org/10.1109/ACCESS.2021.3119569.
Fletz-Brant, C. et al. (2013) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res., 41, W544–W556.
Norouzi, M. et al. (2012) Hamming distance metric learning. In: Advances in Neural Information Processing Systems, vol. 25, MIT Press.
Yun-Tao, Z. et al. (2005) An improved TF-IDF approach for text classification. J. Zhejiang Univ.-Sci. A, 6, 49–55.
Zhang, S. et al. (2022) Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data. Brief. Bioinformatics, 23, 1–10.