Briefings in Bioinformatics, 23(1), 2022, 1–12

https://doi.org/10.1093/bib/bbab445
Problem Solving Protocol

MAResNet: predicting transcription factor binding sites by combining multi-scale bottom-up and top-down attention and residual network

Ke Han†, Long-Chen Shen†, Yi-Heng Zhu, Jian Xu, Jiangning Song and Dong-Jun Yu

Corresponding authors: Dong-Jun Yu, School of Computer Science and Engineering, Nanjing University of Science and Technology, China. E-mail: njyudj@njust.edu.cn; Jiangning Song, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. E-mail: jiangning.song@monash.edu
†These two authors contributed equally to this work.

Abstract

Accurate identification of transcription factor binding sites is of great significance in understanding gene expression, biological development and drug design. Although a variety of methods based on deep-learning models and large-scale data have been developed to predict transcription factor binding sites in DNA sequences, there is room for further improvement in prediction performance. In addition, effective interpretation of deep-learning models is greatly desirable. Here we present MAResNet, a new deep-learning method, for predicting transcription factor binding sites on 690 ChIP-seq datasets. More specifically, MAResNet combines the bottom-up and top-down attention mechanisms and a state-of-the-art feed-forward network (ResNet), which is constructed by stacking attention modules that generate attention-aware features. In particular, the multi-scale attention mechanism is utilized at the first stage to extract rich and representative sequence features. We further discuss the attention-aware features learned from different attention modules in accordance with the changes as the layers go deeper. The features learned by MAResNet are also visualized through the TMAP tool to illustrate that the method can extract the unique characteristics of transcription factor binding sites. The performance of MAResNet is extensively tested on 690 test subsets with an average AUC of 0.927, which is higher than that of the current state-of-the-art

Ke Han received her M.S. degree in computer science from Nanjing University of Science and Technology in 2009. She is currently a PhD candidate in the School of Computer Science and Engineering at Nanjing University of Science and Technology and a member of the Pattern Recognition and Bioinformatics Group. Her research interests include pattern recognition, machine learning and bioinformatics.

Long-Chen Shen received his M.S. degree in software engineering from Nanjing University of Science and Technology in 2021. He will soon be a PhD candidate in the School of Computer Science and Engineering, Nanjing University of Science and Technology and a member of the Pattern Recognition and Bioinformatics Group. His current interests include pattern recognition, data mining and bioinformatics.

Yi-Heng Zhu received his B.S. degree in computer science from Nanjing Institute of Technology in 2015. He is currently a PhD candidate in the School of Computer Science and Engineering at Nanjing University of Science and Technology and a member of the Pattern Recognition and Bioinformatics Group. His research interests include pattern recognition, data mining and bioinformatics.

Jian Xu received the PhD degree from Nanjing University of Science and Technology, on the subject of data mining in 2007. He is currently a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include event mining, log mining and their applications to complex system management and machine learning. He is a member of both China Computer Federation (CCF) and IEEE.

Jiangning Song is an Associate Professor and group leader in the Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia. He is also affiliated with the Monash Centre for Data Science, Faculty of Information Technology, Monash University. His research interests include bioinformatics, computational biomedicine, machine learning, data mining and pattern recognition.

Dong-Jun Yu received the PhD degree from Nanjing University of Science and Technology in 2003. He is currently a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include pattern recognition, machine learning and bioinformatics. He is a senior member of the China Computer Federation (CCF) and a senior member of the China Association of Artificial Intelligence (CAAI).

Submitted: 25 July 2021; **Received (in revised form):** 6 September 2021

methods. Overall, this study provides a new and useful framework for the prediction of transcription factor binding sites by combining the funnel attention modules with the residual network.

Key words: transcription factor binding site; multi-scale bottom-up and top-down attention; deep learning; residual network; sequence analysis

Introduction

In molecular biology, a transcription factor refers to a protein that regulates the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific nucleotide [1, 2]. Transcription factors bind to the DNA, control gene transcription program, and play a major role in a multitude of important cellular processes, such as basal transcription regulation, differential enhancement of transcription and biological development [3–5]. The transcription factor binding site (TFBS) is a DNA fragment binding to a specific transcription factor, which is usually within the range of 4–30 bp [6–8]. Therefore, accurate prediction of TFBSs plays a critical role in characterizing specific functional characteristics of the genome and explaining how highly specific sequence expression program is orchestrated in complex organisms [9–11]. With the development of high-throughput sequencing technology, there is now a large amount of experimental data about high-quality TFBSs, such as TRANSFAC [12], JASPAR [13], etc. However, it is very time consuming and expensive to identify TFBSs through experimental methods. At the same time, how to extract the critical information of TFBS from the existing massive data remains a significant challenge. Hence, researchers have developed numerous calculation-based methods based on massive amounts of data to identify TFBS [6, 14–19].

In early previous years, the researchers proposed many methods to identify TFBS based on traditional machine learning methods. However, because traditional machine learning algorithms rely heavily on manual extraction of features and difficult-to-process large-scale datasets, the methods based on traditional machine learning have the problem of low prediction accuracy on large-scale datasets. For example, Wong et al. [20] proposed kmerHMM method to identify TFBS. The method trained a Hidden Markov Model (HMM) model as the underlying motif representation and then used belief propagations to extract multiple motifs from the HMM. Ghandi et al. [21] proposed a new classifier, gkm-SVM, to predict DNA-binding sites. This classifier used an efficient tree data structure to calculate the kernel matrix, which is twice as accurate as the original gkm-SVM. However, with the rapid accumulation of sequence data, traditional machine learning algorithms cannot meet the current requirements in terms of prediction accuracy and computing speed.

More recently, with the rapid development of deep learning in the field of computer vision [22, 23] and natural language processing [24], an increasing number of studies have successfully applied the cutting-edge deep learning technology to solve many bioinformatics and computational biology problems [25–27]. For the prediction of TFBS, a number of methods based on deep learning have emerged in recent years [14, 15, 18], and accordingly, the prediction accuracy has been further improved compared to traditional machine learning methods. Alipanahi et al. [16] first proposed DeepBind based on a shallow convolutional network to predict the sequence specificities of DNA- and RNA-binding proteins. Zeng et al. [14] further conducted a systematic exploration of CNN architectures for predicting DNA-protein binding and discussed the key parameters in the network. Inspired by the EM algorithm, Luo et al. [15] proposed a novel expectation pooling method, which combined the CNN to predict DNA-protein binding. This method not only improved the prediction performance of TFBS but also explained the model method from statistical methods and deep learning theory. These CNN-based models have achieved significant performance improvements over traditional machine learning methods. However, as the convolution operation is good at extracting local information, it has obvious disadvantages in processing long sequences. To solve this problem, KEGRU [28] combined the word2vec algorithm with the Bidirectional Gated Recurrent Unit (GRU) network to identify TFBS. Combining the different but complementary advantages of CNNs and RNNs, the researchers also designed a hybrid model to predict TFBS, such as DeepSite [29] and DeepTF [30]. Shen et al. [18] have recently proposed a deep transfer learning-based method, termed SAResNet, which combines the self-attention and the residual structure to predict DNA-protein binding from DNA sequences. Although the state-of-the-art method has achieved an impressive prediction performance, it has the following two shortcomings: First, due to the difference in the data volume among different experimental ChIP-seq datasets, there is a room for further improvement in the prediction performance of the model on some smaller datasets. Second, when using the transfer learning method, we can develop a better deep network architecture to extract the high-dimensional features of the binding sites to improve the prediction accuracy of the model.

In this study, we present MAResNet, a novel deep-learning architecture by combining multi-scale bottom-up and top-down attention and residual network, to improve the prediction of TFBSs in DNA sequences. MAResNet is developed based on the stacking of bottom-up and top-down attention modules and residual modules. In the shallow layer, multi-scale attention modules are used to extract more abundant sequence features, while the attention-aware features of different modules change adaptively as layers go deeper. The residual block represents the current advanced deep network structure. Benchmarking experiments show that compared with the current state-of-the-art methods, MAResNet is able to achieve the best predictive performance, with an average AUC of 0.927 on 690 ChIP-seq datasets. An online webserver of MAResNet is implemented and publicly available at http://csbio.njust.edu.cn/bioinf/maresnet/. In addition, the source code of MAResNet is available at https://github.com/csbio-njust-edu/maresnet.

Materials and methods

Benchmark datasets

To fairly evaluate the performance of our proposed model, we used 690 ChIP-seq datasets, which were previously prepared to evaluate many deep learning architectures such as DeepBind [16], CNN-Zeng [14], Expectation-Luo [15] and SAResNet [18] as

Table 1. The sample distribution of benchmark datasets

Dataset	Subsets	Number of positive samples	Number of negative samples	Total number of samples
global datasets	global-TR[a]/global-VL[b] (90%/10%)	2307 290	2307 290	4614 580
	global-TS[c]	400 000	400 000	800 000
A549	TR[d]/VL[e] (80%/20%)	459 740	459 472	919 212
	TS[f]	114 777	115 045	229 822
H1-hESC	TR/VL (80%/20%)	607 774	608 088	1215 862
	TS	152 155	151 841	303 996
HUVEC	TR/VL (80%/20%)	255 931	255 812	511 743
	TS	63 912	64 031	127 943
MCF-7	TR/VL (80%/20%)	433 823	434 368	868 191
	TS	108 801	108 256	217 057

[a,b]global-TR and global-VL represent the training set and validation set of the global datasets, respectively.
[c]global-TS represents the testing set of the global datasets.
[d,e]TR and VL represent the training set and validation set of the corresponding dataset, respectively.
[f]TS represents the testing set of the corresponding dataset.

the benchmark datasets in this study. The 690 ChIP-seq datasets covered the DNA sequences of 91 human cell types bound to 161 unique regulatory factors (generic and sequence-specific factors), some of which were under various treatment conditions. For each of the 690 ChIP-seq datasets, Zeng et al. [14] divided it into a training subset and a corresponding test subset. Based on these datasets, Shen et al. [18] constructed a set of global datasets for the pre-training stage under the premise of ensuring the independence of the test subsets. All the datasets can be downloaded from http://csbio.njust.edu.cn/bioinf/saresnet/.

In addition, due to limited computing resources, we also selected four groups of ChIP-seq datasets from different cell lines as other benchmark datasets, namely A549, H1-hESC, HUVEC and MCF-7, for the search of the best hyper-parameters. Specifically, we combined the training sets and test sets of several typical cell lines based on 690 ChIP-seq datasets, respectively. The 'cd-hit-est-2d' tool [31] was then applied to keep the independence of the test set. These datasets ranged from 600 000 to 1.5 million with balanced positive and negative samples. We divided them into the training data (TR and VL) and testing data (TS) in a ratio of 8:2. Then, we further divided the training data into the training set (TR) and validation set (VL) in a ratio of 8:2. In the experimental analysis part, we evaluated the performance of the proposed model on the four datasets of different cell lines and 690 independent ChIP-seq datasets. The statistical summary of the datasets is provided in Table 1.

Feature representation

The input to MAResNet is a DNA sequence represented by a binary one-hot feature matrix of size $L \times 4$, where L is the length of the DNA sequence (101 bp in this work) and 4 corresponds to the number of base type (A, C, G, T). In one-hot encoding, a value of 1 was assigned to the corresponding base pair position in the input feature matrix and 0 elsewhere, i.e. [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0] and [0, 0, 0, 1]. A missing or invalid base pair in the DNA sequence was assigned a value of -1 in the input feature matrix.

Model architecture and training procedures

In this work, the input DNA sequence was treated as a picture feature with a size of $L \times 1$ and a channel number of 4 (A, C, G, T). Therefore, predicting the binding site of transcription factors can be regarded as a two-class image classification task in computer vision. Some recent studies have shown that image classification algorithms based on deep learning are also suitable for solving this problem [15, 18, 32]. Recent advances in image classification focus on training feed-forward convolutional networks using 'very deep' structures [23, 33, 34]. Various encoder-decoder modules and attention mechanisms are also widely used in computer vision and natural language processing [24, 35, 36].

Bottom-up and top-down attention

It is inefficient to obtain a wider receptive field by stack convolutional networks. Stacking a 'very deep' network will increase the difficulty of model learning and easily cause gradients to disappear or explode. In this study, we referred to the ideas of the encoder-decoder module [33, 36, 37] and attention mechanism [38] in computer vision, and accordingly proposed an attention mechanism based on funnel structure, termed bottom-up and top-down attention. Figure 1B illustrates the detailed structure of the bottom-up and top-down attention module.

We combined the idea of residual learning and proposed the following attention module:

$$\mathbf{x}_{l+1} = (1 + M_l(\mathbf{x}_l)) \circ T_l(\mathbf{x}_l) \tag{1}$$

where $\mathbf{x}_l$ represents the input matrix of the attention module l, $\mathbf{x}_{l+1}$ is the output feature of the attention module l. $T_l(\mathbf{x}_l)$ means the output of the trunk branch, and $M_l(\mathbf{x}_l)$ refers to the output feature of the mask branch, which uses the bottom-up and top-down structure in this model. The bottom-up and top-down structure mimics the feed-forward and feedback attention process [33]. $M_l(\mathbf{x}_l)$is regarded as the control gate for the output of trunk branch, which lies in the range $[0, 1]$. $\circ$ indicates the element-wise product of the two-feature matrix. Considering that stacking attention modules may cause the gradient to disappear, we construct the soft mask branch as identity mapping by referring to residual learning. The trunk branch is used to extract the features of the deep convolutional network, whereas the mask branch works as feature selectors to enhance the recognizable features and suppress the noise from the trunk branch.

Next, we introduce the bottom-up and top-down attention mechanisms with respect to the specific implementation of the pipeline. Figure 2 vividly shows the difference between the soft mask branch and trunk branch, where the thick lines indicate

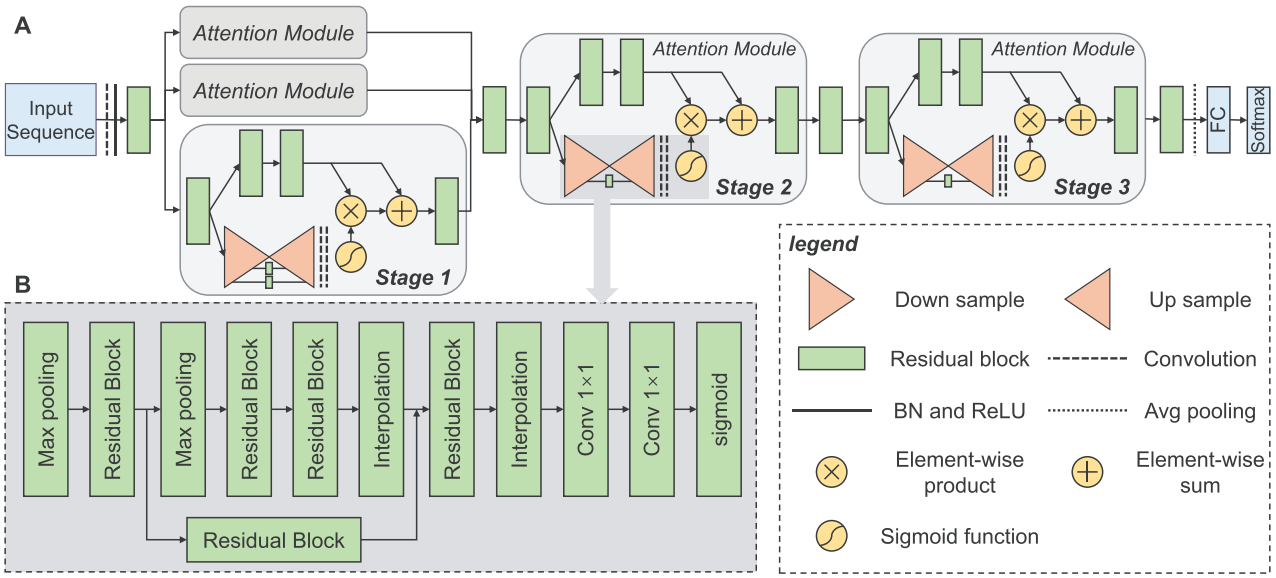

Figure 1. The architecture of MAResNet: (A) an overall architecture of the proposed network. (B) the structure of the bottom-up and top-down attention module.

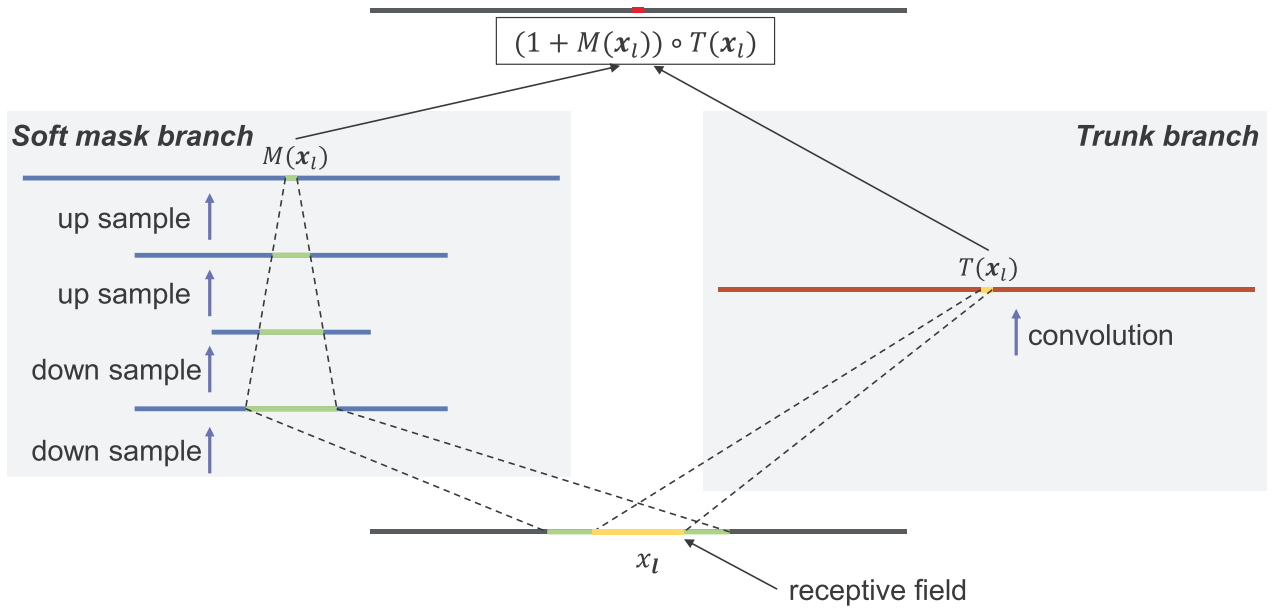

Figure 2. The receptive field comparison between the soft mask branch and trunk branch.

DNA sequence features. The soft mask branch contains a fast feed-forward sweep and top-down feedback step. The bottom-up (down-sample) step quickly collects global information from the sequence features, while the top-down (up-sample) step combines global information with the original feature maps. As shown in Figure 1B, the feature map is sequentially passed through several serial max-pooling layers and residual blocks until the feature map reaches the lowest resolution. Then, the output features containing the global information are expanded through a set of symmetric top-down architectures. In the up-sampling process, bilinear interpolation is used to keep the output feature size the same as that of the input feature map. Further, a sigmoid function normalizes the feature map after two consecutive 1×1 convolutions in a range of [0, 1]. A shortcut connection is also added between the bottom-up and top-down parts to capture the information from different receptive fields.

The pipeline and network architecture of MAResNet

The bottom-up and top-down attention modules can capture more global information and assist the trunk branch to learn more recognizable TFBS features. Based on this, we proposed and implemented our MAResNet pipeline. The architecture of MAResNet is shown in Figure 1. We first used the one-hot encoding to characterize the input DNA sequences ($1 \times L \times 4$). Then, the feature matrix was input to the convolutional layer with 1×3 kernel size, batch normalization layer and ReLU layer successively. During the model implementation, we observed

that the pre-activated residual unit [22] had a better convergence effect than the traditional residual block [23], and thus used the pre-activated unit as the residual block in this work (Supplementary Table S1). Next, we added three attention modules among the four residual blocks, referred to as '*stage 1*', '*stage 2*' and '*stage 3*', respectively. In *stage 1* (shallow layer), we used three different sizes of convolution kernels, namely 3, 5 and 9, to implement the multi-scale attention to focus on the information of different scales of receptive fields. We used 64 convolution kernels with 1×3 kernel size in the attention modules of both *stage 2* and *stage 3*. In each stage, we continuously reduced the size of the feature map, and then used the average pooling layer to further reduce the size of the feature map to reduce the amounts of parameters in fully connected layers (FC). To a certain extent, the application of the average pooling layer can prevent network overfitting and enhance the robustness of the model. Finally, the classification probability was output through a fully connected layer with dropout [39] and the softmax layer was used at the end of the MAResNet architecture.

Model implementation and hyperparameter settings

The model was implemented using the PyTorch framework (v1.8.1) [40] and trained on a single NVIDIA GEFORCE RTX 3090 Graphics Card. In the training process, we utilized the softmax cross-entropy function with the SGD method [41] to optimize the model. The implementation details of the MAResNet model are shown in Table 2. In addition, we adjusted the hyperparameters of the network by comparing the model performance on the validation set of the cell line datasets. The detailed hyperparameter settings are summarized in Table 3. In this study, a grid search was performed on the hyperparameters enumerated in Table 3 on four different cell line datasets (i.e. A549, H1-hESC, HUVEC and MCF-7) to search a set of hyperparameters that can achieve higher accuracy and ensure the execution efficiency of the model. Finally, we applied the searched optimal hyperparameters to model pre-training on the global dataset, and then performed transfer learning on each training subset of the 690 datasets and tested the performance on the corresponding testing subsets.

Performance evaluation metrics

In this study, TFBS prediction is formulated and solved as a binary classification problem. Taking into account the previous studies [18, 29, 30] related to TFBS prediction, we used the accuracy, precision, recall and F1 score as the primary performance measures to evaluate the performance of the developed method. These are defined as follows:

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

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

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

where *TP*, *FN*, *TN* and *FP* denote the numbers of true positives, false negatives, true negatives and false positives, respectively.

Table 2. MAResNet architecture implementation details

Layer	Output Size	Parameters
Conv1d	32×101	1×3, 32, stride 1
Residual block	32×51	$\begin{pmatrix} 1 \times 3, 32 \\ 1 \times 3, 32 \end{pmatrix}$
Attention module stage 1	96×51	Attention-3[a], Attention-5[b], Attention-9[c]
Residual block	64×26	$\begin{pmatrix} 1 \times 3, 64 \\ 1 \times 3, 64 \end{pmatrix}$
Attention module stage 2	64×26	Attention-3[a]
Residual block	64×13	$\begin{pmatrix} 1 \times 3, 64 \\ 1 \times 3, 64 \end{pmatrix}$
Attention module stage 3	64×13	Attention-3[a]
Residual block	64×7	$\begin{pmatrix} 1 \times 3, 64 \\ 1 \times 3, 64 \end{pmatrix}$
Average pooling	64×1	1×7, stride 1
FC, Softmax	2, dropout = 0.8	

[a]Attention-3 indicates that the size of all convolution kernels in the attention module is 1×3.
[b]Attention-5 indicates that the size of all convolution kernels in the attention module is 1×5.
[c]Attention-9 indicates that the size of all convolution kernels in the attention module is 1×9.

However, all these four performance evaluation metrics depend on the prediction cutoff threshold. Therefore, it is important to find rational measures to comprehensively compare different predictors. In this study, the area under the receiver-operating characteristic (ROC) curve (AUC), which is classification-threshold-invariant and reflects the most comprehensive prediction performance, serves as another important evaluation metric [18].

Results and discussion

Funnel attention mechanism improves the prediction performance

To examine the effect of our proposed funnel attention (bottom-up and top-down attention) module, we compared the prediction performance of the networks with and without the funnel attention module on different cell line datasets. To ensure the fairness of the comparison, we made the following modifications based on the baseline. First, to eliminate the influence of multi-scale attention, we modified *stage 1* to a single attention layer with convolution kernel sizes of 3. Second, we constructed a 'non-attention network' by removing the soft mask branch ($M(\mathbf{x}_i) = 0$ mentioned in Figure 2). Finally, we conducted a comparative analysis under the two conditions that the numbers of convolution kernels on *stages 1*, *2* and *3* were 32 and 64, respectively. The performance results of the side-by-side comparison are provided in Table 4.

It can be seen that in most cases, the models with attention modules have consistently improved the prediction performance in terms of various indicators, highlighting the effectiveness and stability of the funnel attention mechanism for the TFBS prediction. In particular, the performance improvement of the models with the '64–64–64' structures were more pronounced than that of the models with the '32–32–32' structures. The effectiveness of funnel attention is more obvious in small-scale networks. In addition, we also found that the performance of the models with the '32–32–32' attention

Table 3. Hyperparameters of MAResNet and the corresponding search space

Calibration parameters	Search space	Sampling	Final settings
learning rate (pre-training)	[0.001, 0.005][a]	fixed-step	0.002
learning rate (transfer learning)	[0.0002, 0.001][b]	fixed-step	0.0004
batch size (pre-training)	{64, 128, 256}	all evaluation	128
batch size (transfer learning)	{32, 64, 128}	all evaluation	64
attention module kernel numbers	{32, 64, 128}	all evaluation	64
optimizer	SGD	fixed	SGD
weight initialization	truncated normal	fixed	truncated normal
dropout ratio	{0.6, 0.7, 0.8}	all evaluation	0.8

[a]step = 1e-3.
[b]step = 2e-4.

Table 4. Performance comparison of the models with different structures on different cell line datasets

Dataset	Structure	32–32–32[a]					64–64–64[b]				
		Accuracy	Precision	Recall	F1 score	AUC	Accuracy	Precision	Recall	F1 score	AUC
A549	non-att[c]	0.817	0.861	0.755	0.805	0.901	0.829	0.875	0.763	0.816	0.912
	with-att[d]	0.828	0.890	0.747	0.812	0.912	0.838	0.879	0.788	0.829	0.918
H1-hESC	non-att	0.807	0.850	0.745	0.795	0.890	0.830	0.811	0.859	0.835	0.916
	with-att	0.820	0.892	0.728	0.802	0.907	0.840	0.846	0.833	0.839	0.919
HUVEC	non-att	0.811	0.798	0.833	0.815	0.898	0.842	0.847	0.833	0.840	0.920
	with-att	0.835	0.904	0.749	0.819	0.920	0.846	0.869	0.813	0.840	0.923
MCF-7	non-att	0.834	0.914	0.739	0.817	0.924	0.843	0.908	0.764	0.830	0.927
	with-att	0.845	0.832	0.865	0.848	0.928	0.853	0.849	0.861	0.855	0.933

[a,b]32–32–32 and 64–64–64 denote that the convolution kernel numbers of stages 1, 2 and 3 are 32 and 64, respectively.
[c]non-att means the model's structure does not contain the attention module.
[d]with-att means the model's structure contains the attention module.

structures was close to or even better than that with the '64–64–64' non-attention structures, on A549, HUVEC and MCF-7 datasets. Among the various performance indicators, we paid more attention to the two comprehensive indicators F1 score and AUC. All the performance indicators except AUC are dependent on the prediction cutoff threshold. Thus, we used AUC to reflect the comprehensive performance of the models. In terms of the AUC indicator, we can see that the performance of the '32–32–32' attention mechanism was improved by 2.4% on the HUVEC dataset compared with '32–32–32' non-attention structure.

The deep network without the attention module is a deep residual network with excellent performance. The residual structure ensures the trainability of the deep convolutional network. The convolutional layer can capture local information, while the global information of DNA sequence can be obtained through deep stacking. However, certain feature information might be lost by stacking the convolutional layer by layer. The attention module we proposed included both down-sampling and up-sampling. The former ensures the acquisition of the global feature information in the feature map, while the latter serves to merge the high-level features with low-level features. This would enable the model to better combine the high-level and low-level features of the context, and add them to the trunk branch through softmax functions and shortcuts (similar to the residual structure). This also enabled the attention layer to play an important role in enhancing important features and suppressing noise. Overall, the experimental results obtained herein suggest that the funnel attention module could indeed improve the prediction performance by adjusting the feature weights of the trunk branch.

Visualization of the attention weights

In this section, we designed a set of comparative experiments to further examine the effectiveness of the multi-scale attention mechanism in feature extraction. Similar to previous experiments, we conducted the experiments on four cell line datasets. Due to the short length of the feature map in the deep attention layer, the effect of multi-scale attention was not significant. We only used the multi-scale attention in stage 1 of the attention module. The experiments indicate that the multi-scale attention with convolution kernel sizes of 1×3, 1×5 and 1×9 (the number of convolution kernels is all 32) achieved a good balance between the model performance and computing power consumption. Figure 3 shows the performance comparison of the multi-scale attention with single-scale attention with convolution kernels of 64 and 96 on these datasets in terms of Accuracy, Precision, Recall, F1 score and AUC. The results demonstrate that multi-scale attention achieved the best performance in models with similar number of parameters across the four different datasets.

Next, to better interpret the multi-scale attention, we partially visualized the attention weights $M(\mathbf{x}_l)$ of the first and second stages based on a positive sample and a negative sample. As shown in Figure 4, the weight of visualization shows that different scales of attention modules focus on different regions. Figure 4A shows the visualization of the attention weights of a positive sample. The weights of the positive sample's attention are approximate on the feature map learned on the small-size attention module, while the model focuses on the middle region on the feature map of the middle- and large-scale attention modules. In the second stage, the model pays more attention to the middle position of the feature map, i.e. the location of the TFBS. As can be seen from Figure 4B, the small- and

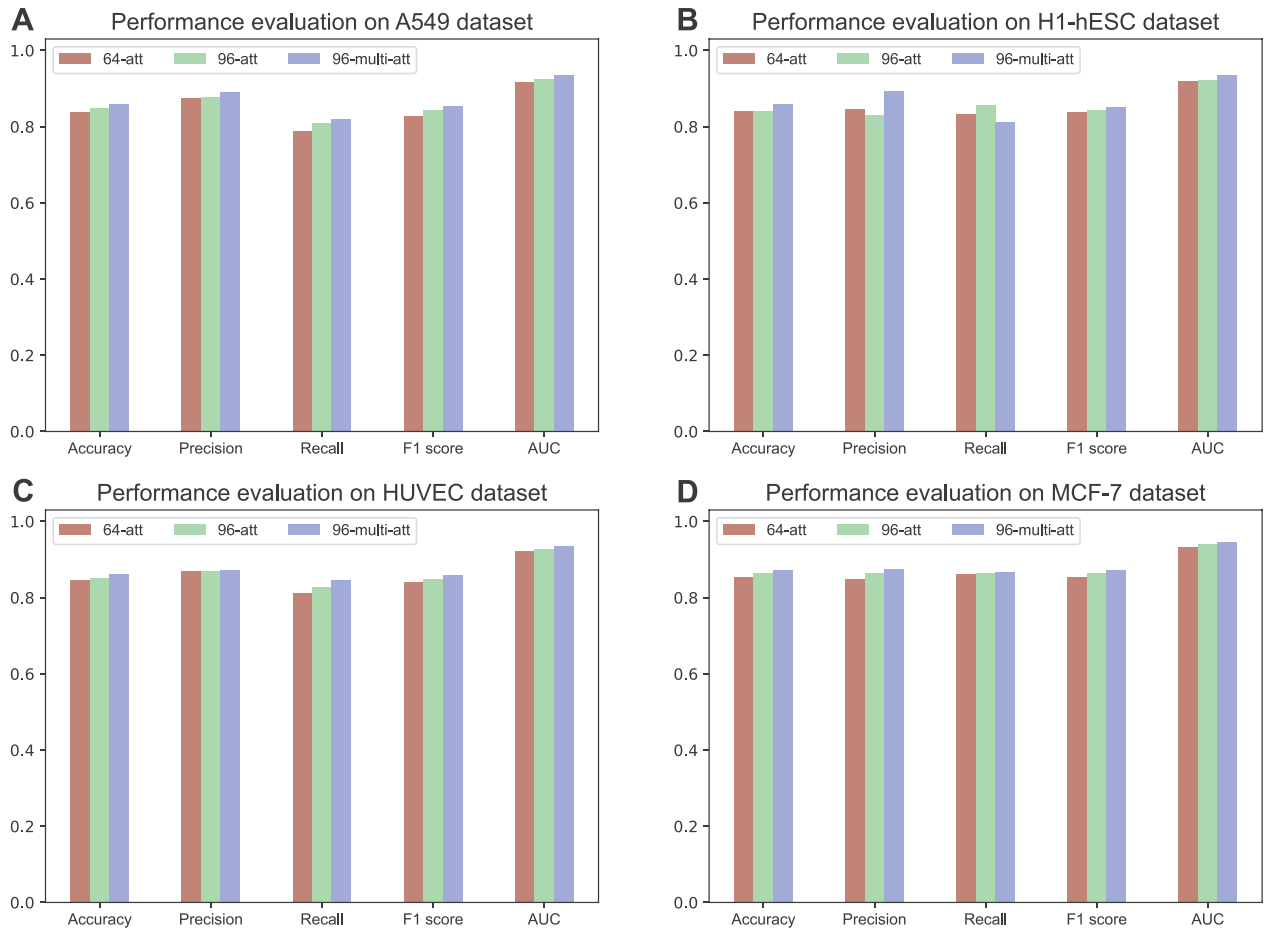

Figure 3. Performance evaluation of models based on multi-scale attention mechanism on A549, H1-hESC, HUVEC and MCF-7 datasets. 64-att refers to the model with 64 1×3 convolution kernels, 96-att refers to the model with 96 1×3 convolution kernels and 96-multi-att refers to a multi-scale attention model with three different sizes of covolution kernels (the number of channels for each attention module is 32, and the total is 96). The number of learning parameters of 64-att, 96-att and 96-multi-att is approximately 0.93, 1.40 and 0.88 million, respectively.

medium-scale attention modules focus on the information in the middle and back ends, while the large-scale attention module focuses on the middle region. There is no obvious focus in the second stage. From the above comparative experiments and visualization, we conclude that the multi-scale attention module can be effectively used to extract more diverse information from the feature matrix and contribute to the classification of whether the sequence contains transcription factors binding sites.

Comparing MAResNet with existing predictors

Most existing studies on the prediction of TFBS are usually based on human ChIP-seq datasets from the ENCODE project. Prior to that, HOCNN [42], KEGRU [28] and DeepRAM [32] used 214, 125 and 83 ChIP-seq datasets, respectively, from the ENCODE project to assess the performance of their respective methods. To ensure the integrity of the experiments and fairly evaluate the performance of these models, we utilized all 690 ChIP-seq datasets to evaluate our method and compare with gkm-SVM [43], DeepBind [16], CNN-Zeng [14], DeepTF [30], Expectation-Luo [15] and SAResNet [18], all of which also used all the 690 ChIP-seq datasets. Using the gkm-SVM R package (https://cran.r-project.org/web/packages/gkmSVM), we trained and tested on each of the 690 ChIP-seq datasets with the default parameters to obtain the performance of the gkm-SVM method. The experimental data of DeepBind and CNN-Zeng were obtained from http://cnn.csail.mit.edu/. The experimental result (in terms of AUC) of DeepTF was provided by its authors. We obtained the source code of Expectation-Luo from https://github.com/gao-lab/ePooling, and then reproduced their experiment locally with the default parameters. In addition, the experimental result of SAResNet method was published by its authors. Figure 5 illustrates the performance of MAResNet on 690 datasets in comparison with gkm-SVM, DeepBind, CNN-Zeng, DeepTF, Expectation-Luo and SAResNet. Overall, we can see that MAResNet performed better than all other methods. Specifically, the median AUC of MAResNet reached 0.931, which was better than the suboptimal method (0.923). In terms of the maximum values of AUC, all the other six methods achieved the maximum values of greater than 0.990, with the only exception of gkm-SVM. In addition, compared with other methods, MAResNet also slightly improved the minimum AUC (Figure 5). Furthermore, both the upper and lower quartiles have been improved, which indicates that our method is superior to the existing methods and has a strong generalization ability.

Moreover, we further analyzed the performance of the models on different scale datasets according to the data volume

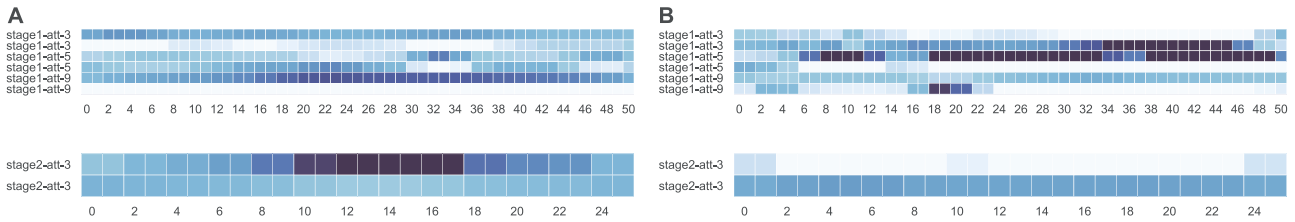

Figure 4. Visualization of the attention weights. Panels A and B represent the attention weights of a positive sample and a negative sample, respectively. The upper two heat maps represent the weight of the first stage of the attention module, and the lower two heat maps represent the weight of the second stage of the attention module. The panels show the weights on part of the attention channels. The darker the square, the more attention a specific weight has received in the output step of that specific layer.

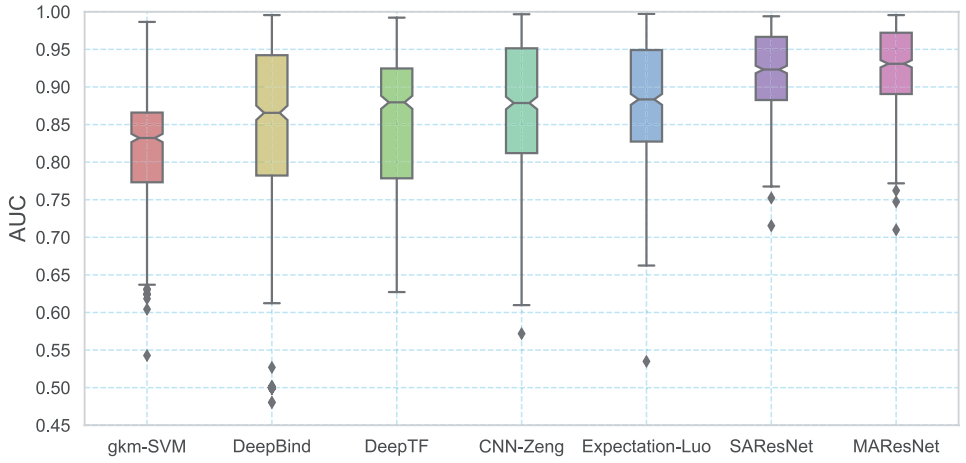

Figure 5. The distribution of AUCs across 690 ChIP-seq datasets for transcription factor binding site prediction. For each box, the intermediate line indicates the median, and the top and bottom edges of the box indicate the upper and lower quartiles, respectively. The upper and lower sides indicate the maximum and minimum values; the diamond marks indicate the outliers.

Table 5. Performance of MAResNet and other existing methods for transcriptional factor binding site prediction on the datasets with different scales

Method	All datasets	Small datasets	Medium datasets	Large datasets	P-value[a]
MAResNet	0.927	0.883	0.914	0.972	-[b]
SAResNet	0.920	0.876	0.907	0.966	1.6×10^{-5}
Expectation-Luo	0.881	0.835	0.859	0.947	3.1×10^{-17}
CNN-Zeng	0.875	0.818	0.850	0.953	1.2×10^{-17}
DeepTF	0.845	0.809	0.818	0.919	2.9×10^{-19}
DeepBind	0.830	0.785	0.809	0.896	7.4×10^{-20}
gkm-SVM	0.818	0.798	0.805	0.856	2.9×10^{-20}

[a]The P-values of the student's t-test for assessing the statistical difference in AUC values between MAResNet and the existing transcription factor binding site predictors.
[b]'-' indicates that the corresponding value does not exist.

division rules proposed by Shen et al. [18]. The specific rule is to divide 690 datasets into three categories: small, medium and large according to the thresholds of 3000 and 30 000. Table 5 shows the average AUC scores of MAResNet and other comparison methods on different scales datasets. It is clear that MAResNet achieved a statistically significant performance improvement in terms of AUC (student's t-test, $P < 1.6 \times 10^{-5}$) compared with other methods. The performance of our proposed MAResNet method has been improved across different-scaled datasets, which suggests the robustness of MAResNet. Moreover, to comprehensively understand the performance of the method, we further compared its performance with that of the other six methods on different-scaled datasets in terms of all five evaluation indices based on bar charts. It is worth noting that the precision of the gkm-SVM method is very high but its recall is very low in Figure 6A. We paid more attention to one of the comprehensive performance metrics, i.e. F1 score, which can better reflect the overall performance of the model than Precision and Recall. As shown in Figure 6, we can more intuitively see that the model is better than other existing models in various evaluation indices.

Visualization of network learning features

In this section, we used TMAP [44] to visually evaluate the classification performance of MAResNet on the identification of TFBSs. TMAP is a visualization method for large and high-dimensional data sets and can represent large-scale and high-dimensional data points as a two-dimensional tree.

To explore and explain the classification performance of MAResNet, we visually analyzed the input feature maps of the fully connected layer in the model using TMAP. In Figure 7, we

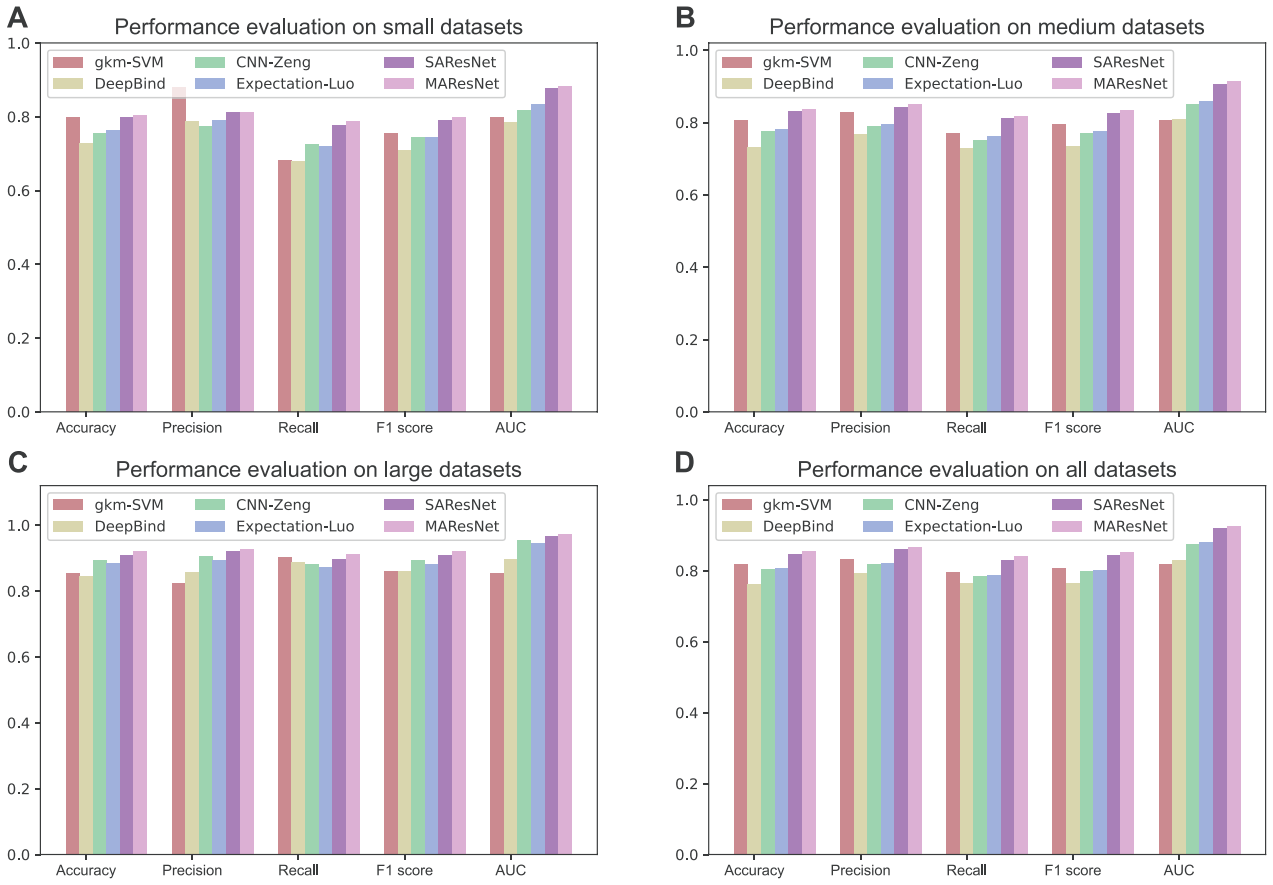

Figure 6. Performance evaluation of the proposed MAResNet method and other existing methods for transcription factor binding site prediction on small, medium, large and all datasets.

randomly selected four different datasets from 690 ChIP-seq datasets for an illustration of data visualization. These four different datasets correspond to specific cell lines and specific TFBSs. For example, 'Dnd41-CTCF' is a ChIP-seq dataset related to T cell leukemia with Notch mutation and CTCF binding factors. From Figure 7, we can observe that the clustering is evident on these four datasets. It is also worth noting that the individual classes are almost completely clustered. This suggests that MAResNet can effectively extract the unique characteristics of TFBSs for sequence-based classification on 690 ChIP-seq datasets.

To further examine the effectiveness of MAResNet, we analyzed the model performance on different cell lines datasets related to the same transcription factor. For example, we trained the model on the 'A549-CTCF' dataset, and then tested the model on other cell line datasets related to the 'CTCF' transcription factor. Although the model was not trained on the corresponding cell line training set, Supplementary Table S2 shows that the model could achieve a very close predictive performance to the model trained on the corresponding cell line training set. These results also demonstrate that MAResNet can effectively learn the characteristic information of the corresponding transcription factors and then accurately identify the binding sites from DNA sequences. Supplementary Tables S3 and S4 also show similar conclusions.

Conclusions

In this work, we have developed MAResNet, a novel deep-learning method for predicting TFBSs in DNA sequences. MAResNet is featured by the fusion of multi-scale bottom-up and top-down attention mechanisms and a state-of-the-art feed-forward network (ResNet). In particular, the network is constructed by stacking funnel attention modules, which generate attention-aware features. In addition, the multi-scale funnel attention is utilized in the first stage. Within each attention module, we also added a shortcut connection between bottom-up and top-down parts to capture the important information from different scales. Benchmarking experiments show that the performance of MAResNet is superior to that of several other existing methods when assessed on the 690 ChIP-seq datasets. The attractive advantages of our network can be reflected in the following three aspects. First, the proposed funnel attention mechanism has been shown to effectively improve the prediction performance by adjusting the feature weights of the trunk branch. Second, comparative experiments and visual analysis indicate that the multi-scale attention module has a capacity to extract more diverse information from the feature matrix, and third, visualization of the features learned by MAResNet through TMAP illustrates that the method can extract the unique characteristics of TFBSs.

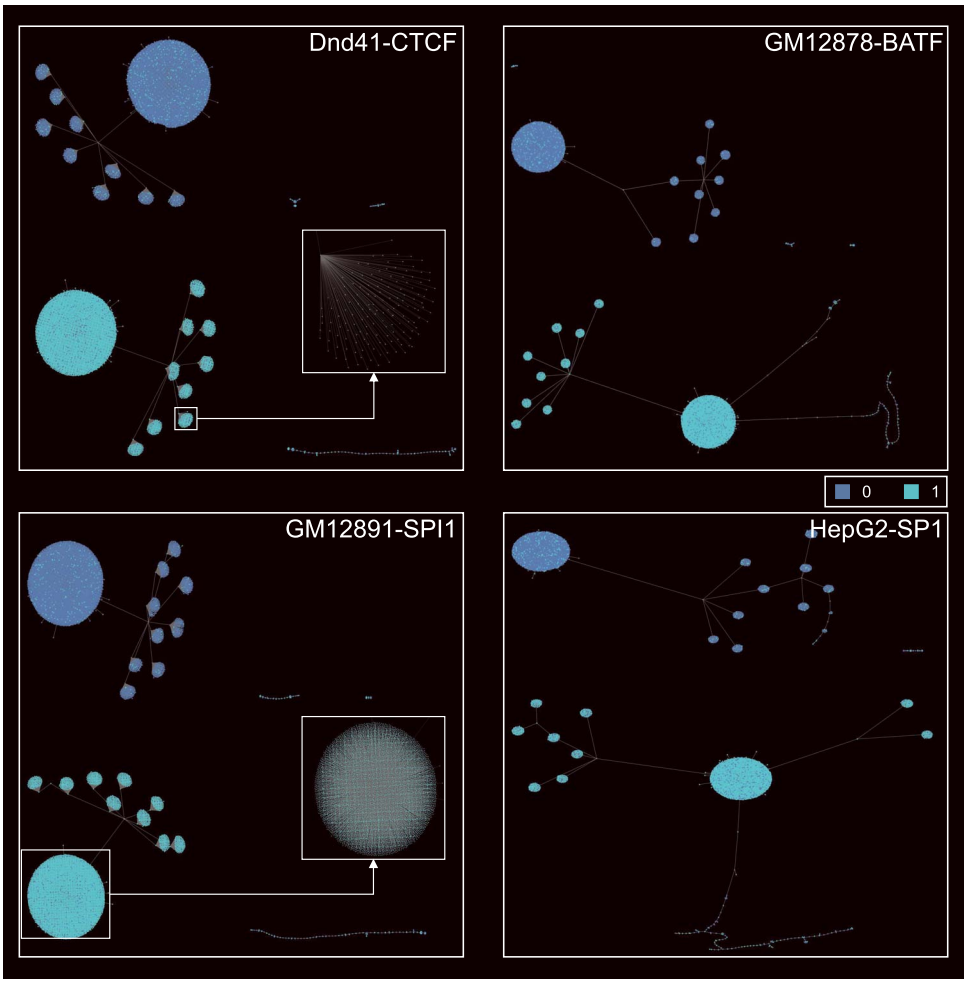

Figure 7. Visualization of the learning features learned by the networks on four different datasets. The darker dots indicate the samples that do not contain the corresponding type of transcription factor binding site, while the lighter dots indicate the samples that contain the corresponding TFBS. The sub-trees in the figure connect the samples that belong to the same category.

Although MAResNet has achieved an excellent performance for predicting TFBSs, there is further room to improvement of MAResNet: Firstly, the soft mask branch may either enhance or weaken the output feature maps of the corresponding trunk branch. To address this, it is possible to design a new activation function to combine the soft mask branch and trunk branch. Secondly, to improve the computing speed of the model, we will develop useful strategies to compress the model while ensuring that the prediction accuracy does not decrease, and finally, it is possible to integrate the funnel attention mechanism with other cutting-edge deep learning architectures to further improve prediction performance. In addition, MAResNet can be generally applied to address other relevant prediction problems in bioinformatics and computational biology [45], such as predicting TFBSs from protein sequences [46–48] and other types of binding and functional sites [49–51]. We believe that MAResNet will greatly help to facilitate our better understanding of deep learning models and the elucidation of gene regulation mechanisms at the genomic sequence level.

Key Points

- This study develops a novel deep-learning method, termed MAResNet, for predicting transcription factor binding sites in DNA sequences.
- MAResNet is featured by the fusion of multi-scale bottom-up and top-down attention mechanisms and a state-of-the-art feed-forward network (ResNet).
- The bottom-up and top-down attention module can improve the prediction performance by enhancing important features and suppressing noise of the trunk branch.
- Benchmarking experiments illustrate that the multi-scale attention module is able to effectively extract more diverse information from the feature matrix and contributes to classifying whether the sequence contains transcription factors binding sites.
- MAResNet achieves a statistically significant performance improvement in terms of AUC compared to

other existing state-of-the-art methods on the 690 ChIP-seq datasets.
- An online webserver of MAResNet (http://csbio.njust.edu.cn/bioinf/maresnet/) is implemented and publicly available for the prediction of transcription factor binding sites, providing a faster tool than existing deep learning-based methods for TFBS prediction.

Availability and Implementation

All the data and source codes used in this study are freely available at http://csbio.njust.edu.cn/bioinf/maresnet or https://github.com/csbio-njust-edu/maresnet.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62072243, 61772273 and 61872186), the Natural Science Foundation of Jiangsu (BK20201304), the Foundation of National Defense Key Laboratory of Science and Technology (JZX7Y202001SY000901), the National Health and Medical Research Council of Australia (NHMRC) (1144652, 1127948), the Australian Research Council (ARC) (LP110200333 and DP120104460), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965) and a Major Inter-Disciplinary Research (IDR) project awarded by Monash University.

References

1. Latchman DS. Transcription factors: an overview. *Int J Biochem Cell Biol* 1997;**29**:1305–12.
2. Karin M. Too many transcription factors: positive and negative interactions. *New Biol* 1990;**2**:126–31.
3. Alexandrov BS, Gelev V, Yoo SW, *et al*. DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. *Nucleic Acids Res* 2010;**38**:1790–5.
4. Li J, J-h O. Differential regulation of hepatitis B virus gene expression by the Sp1 transcription factor. *J Virol* 2001;**75**:8400–6.
5. Wilkinson AC, Nakauchi H, Göttgens B. Mammalian transcription factor networks: recent advances in interrogating biological complexity. *Cell systems* 2017;**5**:319–31.
6. Tompa M, Li N, Bailey TL, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;**23**:137–44.
7. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 2016;**32**:1555–6.
8. Qu K, Wei L, Zou Q. A review of DNA-binding proteins prediction methods. *Current Bioinformatics* 2019;**14**:246–54.
9. Lambert SA, Jolma A, Campitelli LF, *et al*. The human transcription factors. *Cell* 2018;**172**:650–65.
10. Basith S, Manavalan B, Shin TH, *et al*. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput Struct Biotechnol J* 2018;**16**:412–20.
11. Shen Z, Lin Y, Zou Q. Transcription factors–DNA interactions in rice: identification and verification. *Brief Bioinform* 2020;**21**:946–56.
12. Matys V, Kel-Margoulis OV, Fricke E, *et al*. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;**34**:D108–10.
13. Fornes O, Castro-Mondragon JA, Khan A, *et al*. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;**48**:D87–92.
14. Zeng H, Edwards MD, Liu G, *et al*. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* 2016;**32**:i121–7.
15. Luo X, Tu X, Ding Y, *et al*. Expectation pooling: an effective and interpretable pooling method for predicting DNA–protein binding. *Bioinformatics* 2020;**36**:1405–12.
16. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
17. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107–7.
18. Shen L-C, Liu Y, Song J, *et al*. SAResNet: self-attention residual network for predicting DNA-protein binding. *Brief Bioinform* 2021;**22**(5).
19. Zhang Y, Wang Z, Zeng Y, *et al*. High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method. *Brief Bioinform* 2021.
20. Wong K-C, Chan T-M, Peng C, *et al*. DNA motif elucidation using belief propagation. *Nucleic Acids Res* 2013;**41**:e153–3.
21. Ghandi M, Lee D, Mohammad-Noori M, *et al*. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;**10**:e1003711.
22. He K, Zhang X, Ren S, *et al*. Identity mappings in deep residual networks. In: *European conference on computer vision*. Netherlands: Springer, 2016, 630–45.
23. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, USA: IEEE, 2016, 770–8.
24. Devlin J, Chang M-W, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *arXiv preprint arXiv:1810.04805*. 2018.
25. Zhao H, Tu Z, Liu Y, *et al*. PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Res* 2021;**49**:W523–29.
26. Min S, Kim H, Lee B, *et al*. Protein transfer learning improves identification of heat shock protein families. *Plos one* 2021;**16**:e0251865.
27. Liu Y, Zhu Y-H, Song X, *et al*. Why can deep convolutional neural networks improve protein fold recognition? A visual explanation by interpretation. *Brief Bioinform* 2021;**22**.
28. Shen Z, Bao W, Huang D-S. Recurrent neural network for predicting transcription factor binding sites. *Sci Rep* 2018;**8**:15270.
29. Zhang Y, Qiao S, Ji S, *et al*. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. *International Journal of Machine Learning and Cybernetics* 2020;**11**:841–51.

30. Bao X-R, Zhu Y-H, Yu D-J. DeepTF: accurate prediction of transcription factor binding sites by combining multi-scale convolution and long short-term memory neural network. In: *International conference on intelligent science and big data engineering*. Cham: Springer International Publishing, 2019, 126–38.
31. Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
32. Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* 2019;**35**:i269–77.
33. Fe I W, Jiang M, Chen Q *et al*. Residual Attention Network for Image Classification. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, IEEE, 2017, p. 3156–64.
34. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway, NJ, IEEE, 2018, 7132–41.
35. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *Advances in neural information processing systems*, Long Beach, CA, USA, MIT Press, 2017, 5998–6008.
36. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017;**39**: 2481–95.
37. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE international conference on computer vision*. Piscataway, NJ, IEEE, 2015, p. 1520–1528.
38. Wang X, Girshick R, Gupta A *et al*. Non-local neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2018)*. Salt Lake City, USA, 2018, p. 7794–7803. IEEE
39. Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Machine Learn Res* 2014;**15**:1929–58.
40. Paszke A, Gross S, Massa F, *et al*. Pytorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 2019;**32**:8026–37.
41. Bottou L. *Large-scale machine learning with stochastic gradient descent*. Heidelberg: Physica-Verlag HD, 2010, 177–86.
42. Zhang Q, Zhu L, Huang D-S. High-order convolutional neural network architecture for predicting DNA-protein binding sites. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**:1184–92.
43. Ghandi M, Mohammad-Noori M, Ghareghani N, *et al*. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* 2016;**32**:2205–7.
44. Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Chem* 2020;**12**:1–13.
45. Xu L, Jiang S, Wu J, *et al*. An in silico approach to identification, categorization and prediction of nucleic acid binding proteins. *Brief Bioinform* 2021;**22**:bbaa171.
46. Manavalan B, Basith S, Shin TH, *et al*. 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cell* 2019;**8**:1332.
47. Chen W, Lv H, Nie F, *et al*. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 2019;**35**:2796–800.
48. Xu R, Zhou J, Wang H, *et al*. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. In: *BMC systems biology*. London, England: BioMed Central,, 2015, 1–12.
49. Hu J, Li Y, Zhang Y, *et al*. ATPbind: accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons. *J Chem Inf Model* 2018;**58**:501–10.
50. Feehan R, Franklin MW, Slusky JS. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nat Commun* 2021;**12**:1–11.
51. Song J, Li F, Takemoto K, *et al*. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J Theor Biol* 2018;**443**:125–37.