

Locating transcription factor binding sites by fully convolutional neural network

Qinhu Zhang^{ID}, Siguo Wang, Zhanheng Chen, Ying He^{ID}, Qi Liu and De-Shuang Huang

Corresponding authors: Qi Liu, Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Siping Road 1239, Shanghai 200092, China. E-mail: qiliu@tongji.edu.cn; De-Shuang Huang, Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China. E-mail: dshuang@tongji.edu.cn.

Abstract

Transcription factors (TFs) play an important role in regulating gene expression, thus identification of the regions bound by them has become a fundamental step for molecular and cellular biology. In recent years, an increasing number of deep learning (DL) based methods have been proposed for predicting TF binding sites (TFBSs) and achieved impressive prediction performance. However, these methods mainly focus on predicting the sequence specificity of TF-DNA binding, which is equivalent to a sequence-level binary classification task, and fail to identify motifs and TFBSs accurately. In this paper, we developed a fully convolutional network coupled with global average pooling (FCNA), which by contrast is equivalent to a nucleotide-level binary classification task, to roughly locate TFBSs and accurately identify motifs. Experimental results on human ChIP-seq datasets show that FCNA outperforms other competing methods significantly. Besides, we find that the regions located by FCNA can be used by motif discovery tools to further refine the prediction performance. Furthermore, we observe that FCNA can accurately identify TF-DNA binding motifs across different cell lines and infer indirect TF-DNA bindings.

Key words: fully Convolutional Neural Network; global Average Pooling; nucleotide-level prediction; TFBSs location

Introduction

Transcription factors (TFs) can activate or suppress transcription of genes by binding to specific DNA noncoding regions, thereby playing an integral role in gene expression. Previous studies have confirmed that TF binding sites (TFBSs) are some short DNA sequences and relatively conserved in the long-term evolution [1], and generally have specific patterns that are commonly called TF-DNA binding motifs. Identification of TFBSs and their corresponding motifs have become a fundamental step for molecular and cellular biology [2].

Due to the fast development of high-throughput sequencing technology in the last decades, particularly, Chromatin Immunoprecipitation sequencing (ChIP-seq) [3] provides a large amount of TF-DNA binding data and enables new insights into gene regulation. Abundant TF-DNA binding data provide an unprecedented opportunity for developing computational methods to predict TFBSs and motifs. Based on these binding data, a series of computational methods have been proposed for predicting motifs. For example, MEME (Multiple EM for Motif Elicitation) [4], based on expectation maximization (EM), predicted TF-DNA

Qinhu Zhang is now working at Tongji University as a post-doctor at the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University, Shanghai, China.

Siguo Wang is working toward the PhD degree in computer science and technology, Tongji University, China.

Zhanheng Chen is currently a scientific assistant researcher. He received his PhD degree from the University of Chinese Academy of Sciences, China.

Ying He is pursuing a PhD degree in computer science and technology at Tongji University, China.

Qi Liu is a professor at the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, China.

De-Shuang Huang is a chaired professor at Tongji University, China. At present, he is the Director of the Institute of Machines Learning and Systems Biology, Tongji University, China. Dr Huang is currently IAPR Fellow and IEEE Fellow.

Submitted: 13 November 2020; **Received (in revised form):** 11 December 2020

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

binding motifs by searching for repeated, ungapped sequence patterns that occur in the biological sequences. DREME (Discriminative Regular Expression Motif Elicitation) [5] used a simpler, nonprobabilistic model (regular expressions) to describe the short binding motifs characteristic of single TFs, which is often used as the complement of MEME. MEME-ChIP [6] identified motifs from ChIP-seq peak regions by assembling two complementary motif discovery tools: MEME and DREME. However, the high computational complexity of these motif discovery tools restricts the number of input sequences or the range of search space, which may sacrifice the accuracy of identifying motifs. Over the past 5 years, deep learning (DL) have achieved impressive performance in many fields, such as computer vision and natural language processing, inspiring researchers to design DL-based methods to predict TFBSs and motifs [7–9]. For example, DeepBind [10], one of the earliest and most well-verified DL-based algorithms, applied convolutional neural networks (CNNs) to predict the sequence specificity of TF-DNA binding. DeepSea [11], another impressive DL-based algorithm, also used deep CNN to predict TF-DNA binding motifs and the chromatin effects of sequence alterations from large-scale chromatin-profiling data. DanQ [12] predicted TF-DNA binding motifs and prioritized functional SNPs by combining CNN with recurrent neural networks (RNNs). However, these DL-based methods mainly focus on predicting the sequence specificity of TF-DNA binding, and fail to identify motifs and TFBSs accurately. Besides, they view motif discovery as a sequence-level binary classification task, thereby they need to carefully select negative sequences for positive sequences (peak regions), and different selection strategies will give rise to diverse predictions.

In this paper, we developed a novel motif discovery method which is mainly based on a fully CNN coupled with global average pooling, namely fully convolutional network coupled with global average pooling (FCNA). The proposed model FCNA views motif discovery as a nucleotide-level binary classification task, which can (i) avoid generating negative sequences, and (ii) locate some short regions that contain TFBSs, and (iii) predict TF-DNA binding motifs accurately. Specifically, (i) high-quality position counting matrices (PCMs) were collected from the HOCOMOCO motif database [13], by which each nucleotide in DNA sequences was annotated; (ii) FCNA, which incorporates a fully CNN, a global average pooling, and a hard negative mining loss, was trained on the annotated TF-DNA binding data; (iii) the trained FCNA was used to locate TFBSs and predict motifs on the test data. Experimental results on the ChIP-seq datasets show that FCNA outperforms other competing methods significantly. Besides, FCNA was first to locate some short regions that contain TFBSs, on which motif discovery tools were then trained to predict TF-DNA binding motifs. As a result, we find that the regions located by FCNA can contribute to further refining the performance of predicting motifs. Furthermore, according to the predicted motifs, we observe that FCNA can accurately identify TF-DNA binding motifs across different cell lines and infer indirect TF-DNA bindings.

Materials and methods

Data preparation

We collected 53 TF ChIP-seq datasets from the ENCODE project, which are separately from three cell lines including A549 (20), GM12878 (21), and MCF7 (12), and downloaded high-quality PCMs (marked as A) from the HOCOMOCO motif database. For each TF dataset, 500 bp regions surrounding peaks were extracted, and

its corresponding PCM was used to annotate each nucleotide in the 500 bp regions as 0 or 1 in which label '1' means that the nucleotide belongs to TFBSs. Briefly, since PCM not only provide the counting number of four nucleotides at each position but also the exact length of the corresponding motif, each region of the same length as the motif was scored by the PCM, and then the region with the highest score was chosen as the positive data, which assumes that the chosen region is the TFBS. As we known, TFBSs are short sequences ranging from 5 bp to 22 bp. Therefore, the annotated data are extremely imbalanced in the experiments, and the ratio of negative to positive is about 32.

The framework of FCNA

The fully convolutional network (FCN) was originally applied to image segmentation [14], which replaces all fully-connected layers with convolutional layers and can take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. FCN often adapts classification networks (e.g. VGG [15], ResNet [16]) into FCN, and uses deconvolution operations to gradually restore downsampled feature maps to the original image size, and uses a skip line to combine semantic information from a deep layer with appearance information from a shallow layer. Another typical segmentation model U-Net [17] was applied to biomedical image segmentation, which is similar to FCN but adopts a symmetrical architecture, and uses upsample operations and normal convolution operations to gradually restore downsampled feature maps to the original image size.

Inspired by the original FCN and U-Net, we designed a novel motif discovery method namely FCNA for locating TFBSs and predicting TF-DNA binding motifs. As shown in Figure 1, FCNA is a symmetrical architecture, which consists of a top-down encoding process (left), a bottom-up decoding process (right). The source code and data are available at: <https://github.com/turningpoint1988/FCNA>.

Top-down encoding process

This process contains three convolutional blocks and a global average layer, in which each block is composed of a convolutional layer, a ReLU layer, a max-pooling layer and a dropout layer. The computation process can be mathematically described in Equation (1).

$$\begin{aligned} X &= \max(0, S \odot M + b) \\ Y &= \text{maxpool}(X) \\ Z &= \text{dropout}(Y) \end{aligned} \quad (1)$$

where S is a one-hot matrix encoded by the one-hot method, and X means the convolution of the convolutional kernel M and the matrix S , and \odot denotes the convolution operation. The convolutional layer is often viewed as motif scanners, which is used to score each segment of sequences. The max-pooling layer is used to reduce the computational complexity and select the best response of local adjacent regions. The dropout layer is often used to alleviate the overfitting problem.

In the application of image segmentation, several studies have demonstrated that global average pooling can capture the global context of images [18, 19]. Similarly, the global context of DNA sequences is also important for motif discovery, e.g. the information of flanking regions can influence the binding activity of TF-DNA [20]. Hence, we added a global average pooling layer on the top of the last convolutional block to capture the global context of TF-DNA binding sequences.

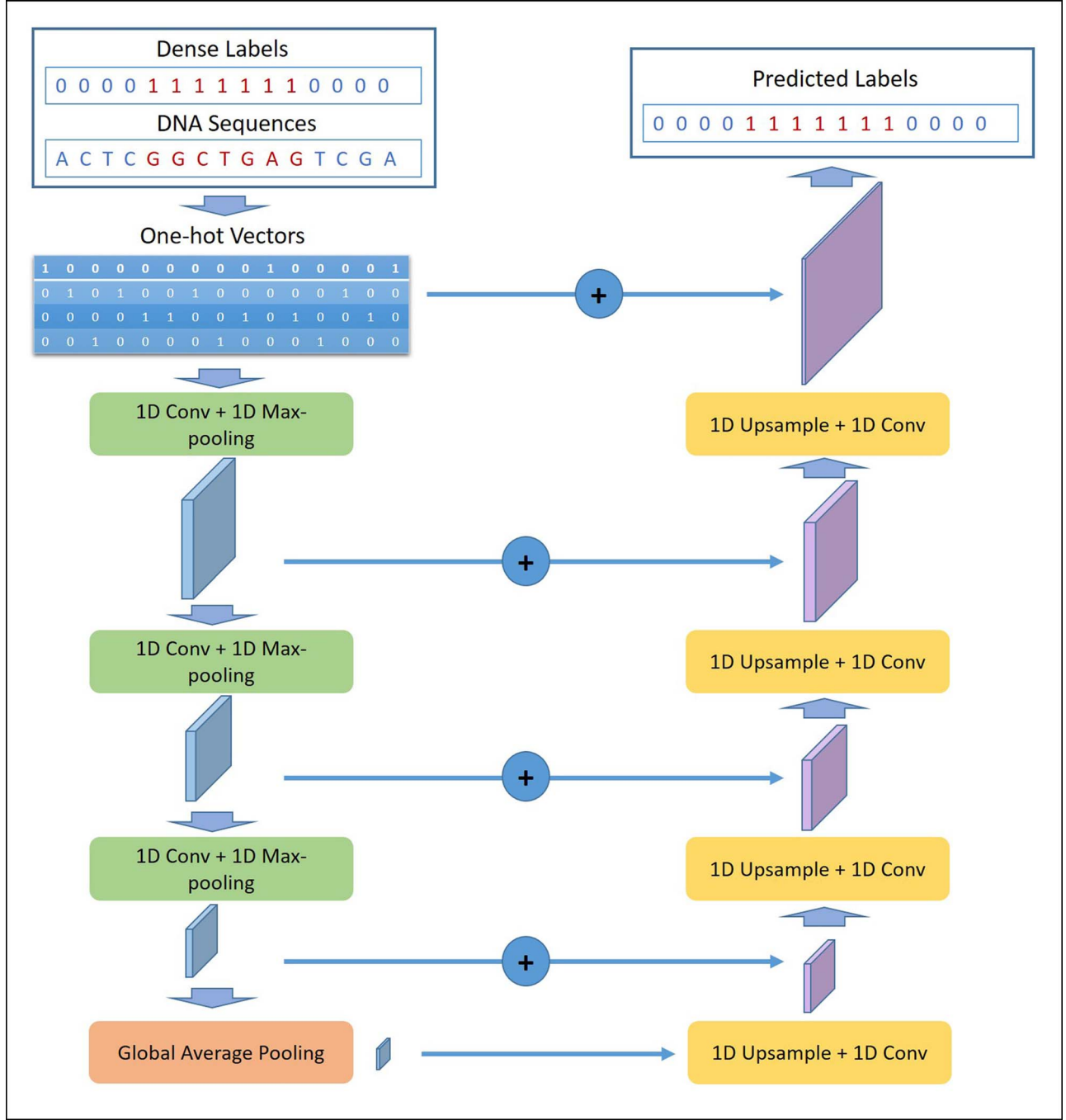


Figure 1. The framework of FCNA, which mainly contains a top-down encoding process (left), a bottom-up decoding process (right).

Bottom-up decoding process

This process, symmetrically, has four deconvolutional blocks and four skip lines, in which each block is composed of an upsample layer, a batch normalization (BN) layer, a ReLU layer, and a convolutional layer, and each line is a summation operation. The computation process can be mathematically described in Equation (2).

$$\begin{aligned}
 Y^+ &= \text{upsample}(Z^+) \\
 Y^+ &= Y^+ + Z \\
 Y^+ &= \text{BN}(Y^+) \\
 X^+ &= \max(0, Y^+) \\
 S^+ &= X^+ \odot M^+ + b^+
 \end{aligned}
 \tag{2}$$

where Z is the outputs of the encoding process at the same level and \odot denotes the convolution operation. The upsample layer is used to restore the size of downsampled feature maps. The skip line is used to combine high-level semantic information from a deep layer with detailed location information from a shallow layer, which contributes to accurately predicting the label of each position.

Hard negative mining loss

Since the annotated data are extremely imbalanced, the normal binary classification is not suitable for this situation. However, in the field of object detection, a few efficient methods have been

developed to solve the problem of imbalanced data. Particularly, the hard negative mining method [21] is commonly-used to deal with imbalanced positive and negative data. Following the concept of the method, we designed a hard negative mining loss for locating TFBSs and predicting motifs, which is briefly described as follows: (i) computing the losses of all positive and negative data; (ii) sorting the losses of the negative data, and selecting the top-k losses of them where k is determined by a specified ratio; (iii) separately taking the average of the losses of all the positive data and the losses of the selected negative data, and then outputting the summation of them. The computation process can be mathematically described in Equation (3).

$$\begin{aligned} Loss_{pos} &= \text{Crossentropy}(S_{pos}^+) \\ Loss_{neg} &= \text{Crossentropy}(S_{neg}^+) \\ Loss_{neg}^+ &= \text{top-k}(Loss_{neg}, \text{ratio} = 0.3) \\ Loss &= \text{mean}(Loss_{pos}) + \text{mean}(Loss_{neg}^+) \end{aligned} \quad (3)$$

where S^+ is the output of the last deconvolutional block, and the ratio was set to 0.3 in this paper.

Locating TFBSs and predicting motifs

As shown in Figure 1, since the outputs of FCNA are the nucleotide-level predictions, so a post processing of them is needed to locate TFBSs and to predict motifs.

Locating TFBSs

Firstly, a threshold value (e.g. 0.9) was set empirically, and the probabilities of all nucleotides were transformed into 1 if bigger than the threshold value or 0 otherwise. Secondly, a sliding window of short length 50 bp (the motif length $< 50 \text{ bp} < \text{sequence length}$) was used to count the number of label 1. Thirdly, the region containing the maximum number of label 1 was selected from the whole sequence as the location of TFBSs. Since it is difficult to precisely locate TFBSs with 100%, we can use a window of short length to roughly locate them. In the later experiments, we will show that the located regions can be used to further refine the prediction performance.

Predicting motifs

Firstly, the process of locating TFBSs described above was repeated to obtain the located regions from DNA sequences. Secondly, the trained weights of the first convolutional layer were used to score each subregion of the located regions, from which the ones with the highest score were then selected. Thirdly, these selected subregions were aligned to compute position frequency matrixes (PFMs). Finally, TOMTOM [22] was utilized to match the PFMs with experimentally validated motifs from standard databases.

Evaluation metrics

In this paper, Intersection over Union (IOU) was used to test the nucleotide-level prediction performance of FCNA, which is a commonly-used metric in the field of image segmentation. IOU can measure the overlapping ratio of the predicted labels and the true labels, and gets to 1 if completely overlap. To test the performance of predicting motifs, three statistical significances including P-value, e-value and q-value were adopted, which are commonly-used for comparing predicted motifs with experimentally validated motifs [19].

In the process of evaluation, a 5-fold cross-validation strategy was employed. In other words, all data are randomly divided into five parts, and four parts of them are used as the training data, whereas the remaining one is used as the test data.

The competing methods

In this paper, we designed three comparative methods including CNN, CNN+ and MEME-ChIP.

CNN: its architecture is similar to DeepSea, which has three convolutional layers. Here we did not compare DeepBind, as our previous studies have shown that the deep model (DeepSea) is better than the shallow model (DeepBind) in the task of predicting TFBSs [23, 24].

CNN+: its architecture is similar to DeepSea, but it has prior knowledge about the length of motifs, which means that the kernel size of the first convolutional layer is the same as the length of motifs. Here we wanted to investigate the effect of the motif length on the prediction performance.

MEME-ChIP: it integrates two complementary motif discovery tools: MEME and DREME, which are the most commonly-used traditional methods.

Experimental results

Hyperparameters selection

In this section, some important hyperparameters were investigated by using 20 TF-DNA binding datasets from A549, whereas other hyperparameters were set to default values. The evaluation metric is mean IOU (miou).

As shown in Figure 2A, the effect of the hard negative mining loss on the prediction performance of FCNA was investigated. In this experiment, we manually set a specified value (e.g. 0.3) to select hard negative examples (note that the specified value 0.3 was not validated). From the results, we can see that FCNA with the hard negative mining loss significantly outperforms FCNA with the normal loss, and the performance gain is 0.095, which demonstrates that the hard negative mining loss is very efficient for handling imbalanced data.

As shown in Figure 2B, the effect of global average pooling on the prediction performance of FCNA was the effect of the hard negative mining loss on the prediction performance of FCNA was investigated. In this experiment, both FCN and FCNA adopted the hard negative mining loss and other same settings except global average pooling. From the results, we can find that FCNA significantly outperforms FCN, and the performance gain is 0.096, which demonstrates that global average pooling is very efficient for locating TFBSs, and also shows that the global context of TF-DNA binding sequences is very important.

As shown in Figure 2C, the effect of the threshold value on the prediction performance of FCNA was the effect of the hard negative mining loss on the prediction performance of FCNA was investigated. In this experiment, we set three threshold values {0.5, 0.7, and 0.9} to transform the probabilities of outputs into 1 or 0. From the results, we observe that the performance of using 0.7 is better than that of using the other two values, and the performance gains are 0.01 and 0.02. Intuitively, we are more concerned about the true-positive samples, so a big value was set to decrease some false-positive samples, but it may filter out too many true-positive samples if too big. The detailed results were recorded in Supplementary Table 3.

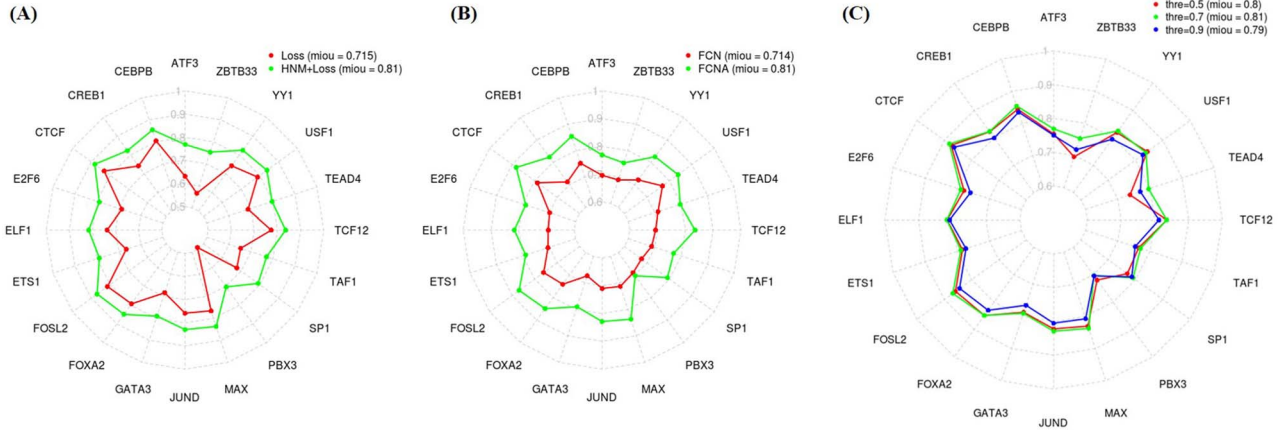


Figure 2. The performance comparison of three important hyperparameters, including hard negative mining loss (A), global average pooling (B), different threshold values (C).

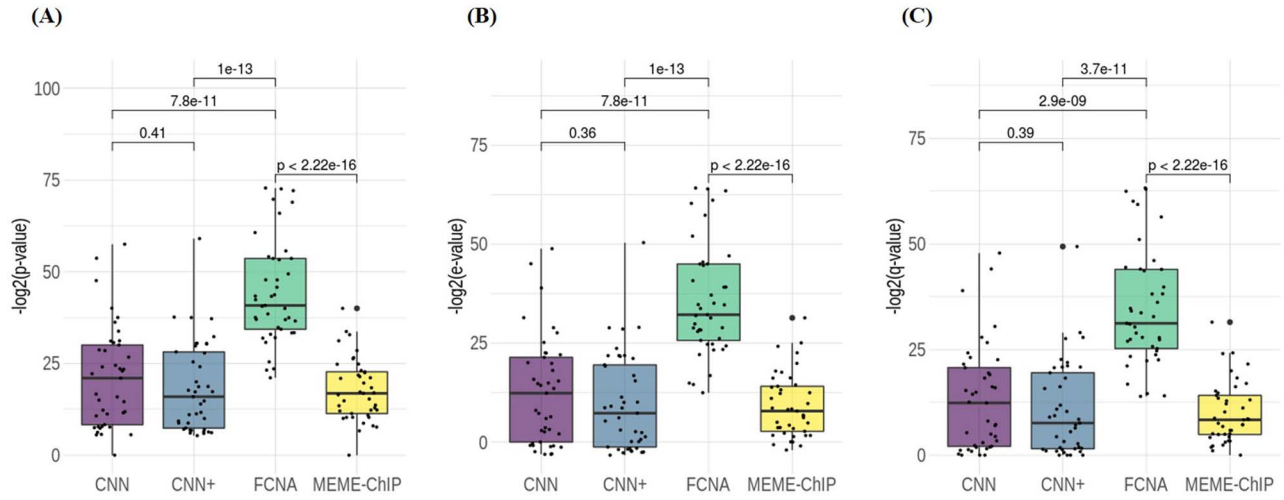


Figure 3. The performance comparison of four motif discovery methods. The adopted metrics are separately $-\log_2(P\text{-value})$ (A), $-\log_2(e\text{-value})$ (B) and $-\log_2(q\text{-value})$ (C).

The performance of predicting motifs

In this section, we first made similarity comparisons between motifs predicted by FCNA from 41 TF-DNA binding datasets (A549 + GM12878) and experimentally validated motifs in the HOCOMOCO database by using TOMTOM. Then, similar comparisons were extended to the other three competing methods including MEME-ChIP, CNN and CNN+. The evaluation metrics are $-\log_2(P\text{-value})$, $-\log_2(e\text{-value})$ and $-\log_2(q\text{-value})$.

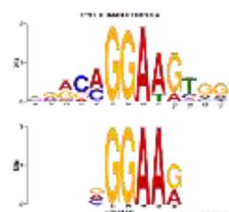
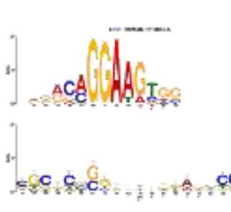
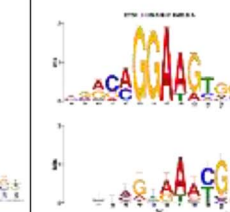
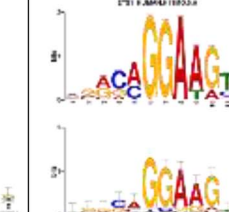
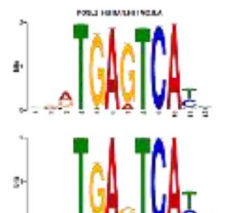
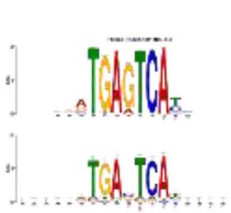
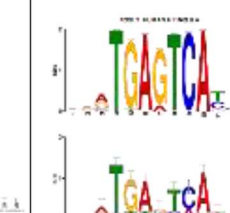
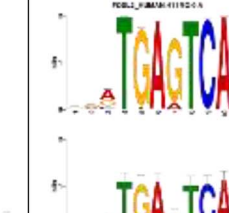
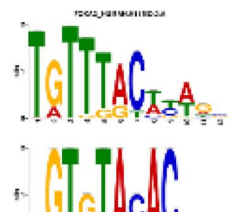
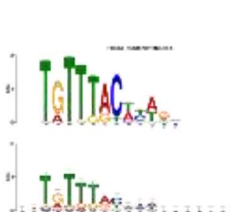
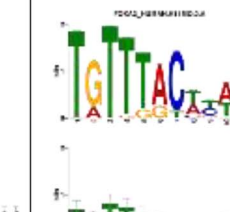
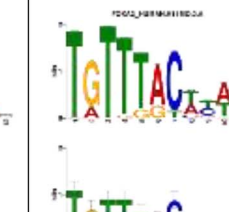
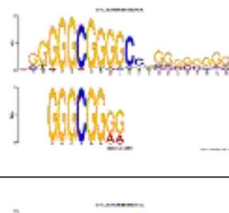
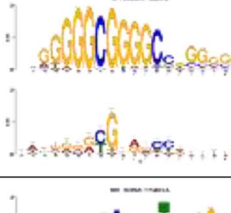
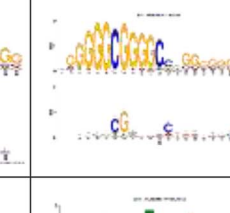
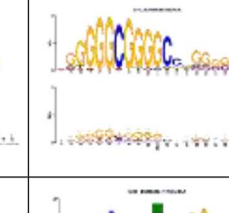
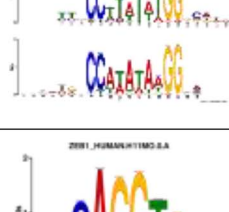
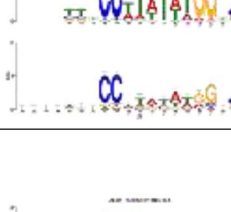
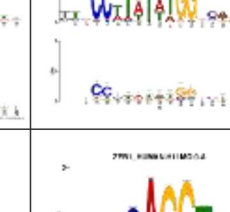
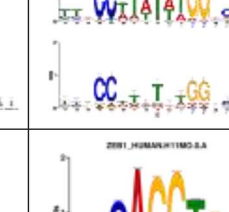
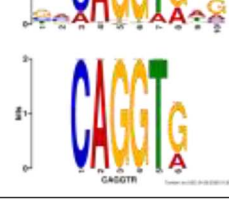
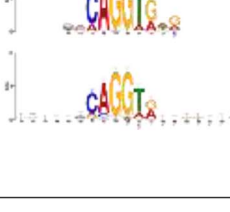
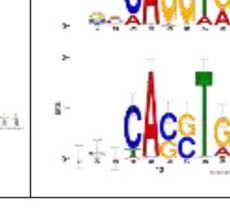
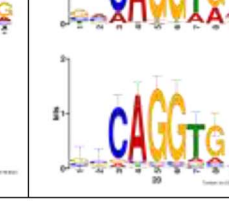
First, we compared CNN with CNN+ to investigate the effect of the prior knowledge (motif length) on the performance of predicting motifs. As shown in Figure 3, the $-\log_2(P\text{-value})$, $-\log_2(e\text{-value})$ and $-\log_2(q\text{-value})$ of CNN are slightly higher than the values for CNN+ (Wilcoxon test P-values are 0.41, 0.36 and 0.39, respectively). Furthermore, the sequence-level prediction performance between them was compared, and the evaluation metrics AUC and PRAUC were used. As shown in Supplementary Figure 1, the sequence-level prediction performance of CNN and CNN+ is almost the same. The above results demonstrate that the motif length cannot significantly influence the final prediction performance.

Then, we compared FCNA with CNN, CNN+ and MEME-ChIP to test the motif prediction performance of FCNA. As shown in

Figure 3, the $-\log_2(P\text{-value})$, $-\log_2(e\text{-value})$, and $-\log_2(q\text{-value})$ of CNN are significantly higher than the values for other competing methods (Wilcoxon test P-values are $<2.9e-09$, $<3.7e-11$ and $<2.22e-16$, respectively), which quantitatively demonstrates that FCNA is much better than others in the task of predicting motifs. As shown in Table 1 and Supplementary Table 1, all the motif logos found by FCNA are better matched with the motif logos in the HOCOMOCO database than other methods, which visually shows that FCNA is superior to others. The above results confirm that the proposed method FCNA is very efficient for predicting motifs. The detailed results were recorded in Supplementary Table 4.

Finally, we explored the ability of FCNA to find indirect TF-DNA binding motifs. Following the process of predicting motifs, TOMTOM was used to match some validated motifs with high $-\log_2(P\text{-value})$. Except for the target TF (which generally has the highest $-\log_2(P\text{-value})$), the top-5 TFs from all the matched TFs were picked out according to the $-\log_2(P\text{-value})$. As shown in Figure 4 and Supplementary Table 5, we observe two scenarios: (i) the matched TFs with high $-\log_2(P\text{-value})$ may interact with the target TF, and cobind to neighboring sites; (ii) the matched TFs and the target TF belongs to the same TF family. For the

Table 1. Motif logos comparison of different methods

TF(A549)	MEME-CHIP	CNN	CNN+	FCNA
ETS1				
FOSL2				
FOXA2				
TF(GM12878)	MEME-CHIP	CNN	CNN+	FCNA
SP1				
SRF				
ZEB1				

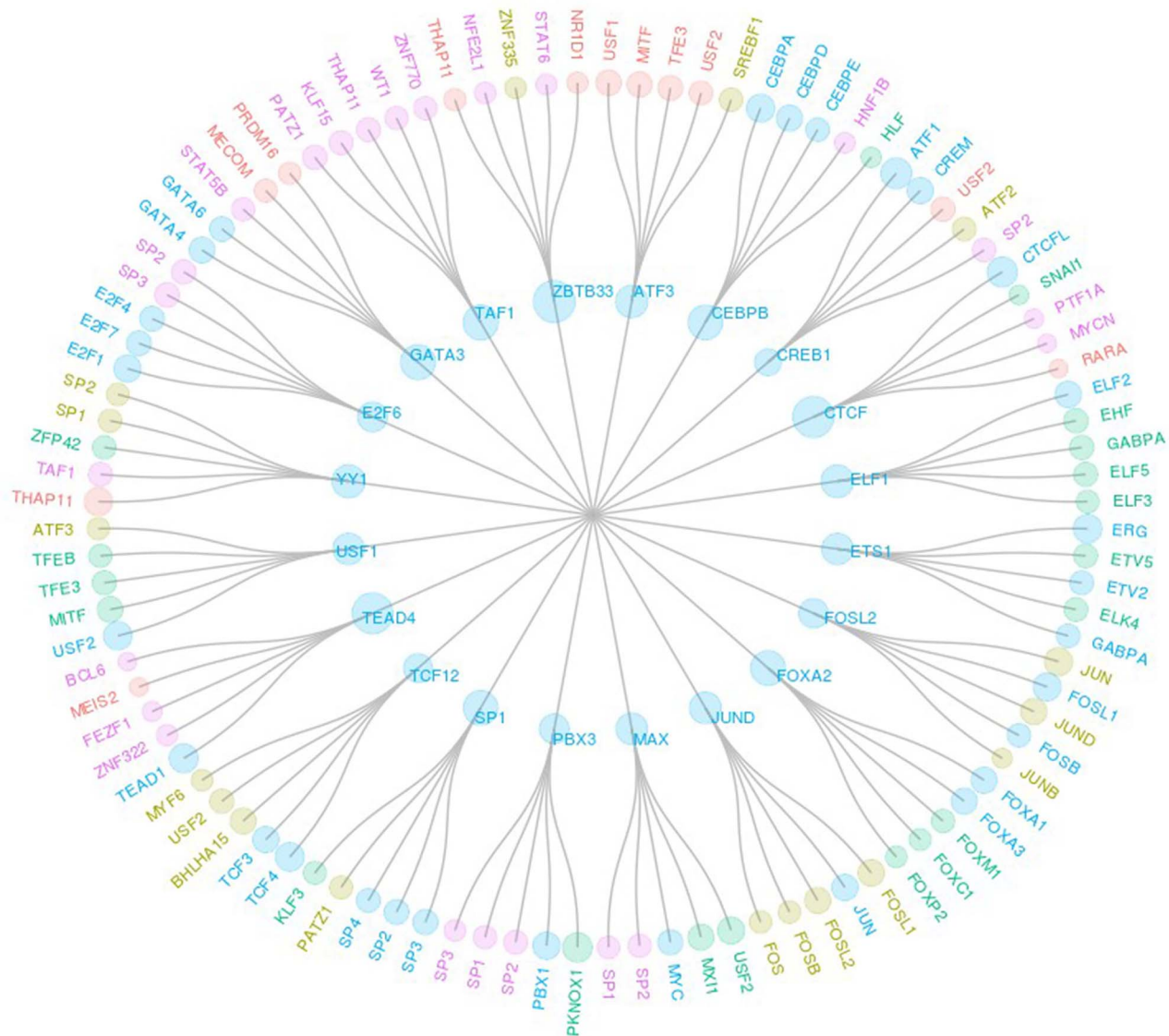


Figure 4. The TFs found by FCNA. The TFs of the inner loop are the target TFs and the TFs of the outer loop are the indirect TFs, where the size of circle corresponds to the $-\log_2(\text{P-value})$ value. Different colors are used to designate different TF classes, where brown means ‘super class’, and yellow means ‘class’, and green means ‘family’, and blue means ‘sub-family’, and pink means ‘other super classes’. The TFs marked by green and blue belong to the same TF family sharing the consensus binding sequence while the ones marked by other colors are more likely to cobind with the target TF.

first scenario, for example, the matched TFs of ATF3 contain USF1 and USF2, where ATF3 belongs to the basic leucine zipper family whereas USF1 and USF2 belong to the basic helix-loop-helix leucine zipper family, but ATF3 has a tethered binding with USF (USF is the heterodimer of USF1 and USF2) [25]. The matched TFs of FOSL2 contain JUN-related TFs (JUN, JUND and JUNB) meanwhile the matched TFs of JUND also contain FOS-related TFs (FOSL1, FOSL2, FOSB and FOS), and the related study has demonstrated that all JUN-FOS heterodimers strongly bind to the TPA-response element [26]. For the second scenario, for instance, the matched TFs of SP1 contain SP-related TFs (SP2, SP3 and SP4), and the matched TFs of CEBPB contain CEBP-related TFs (CEBPA, CEBPD and CEBPE), and the matched TFs of CTCF contain CTCFL, as they belong to the same TF family and share the consensus binding sequence. According to the matched TFs of all target TFs, we find that in most cases, majority of the matched TFs have the same TF family as the

target TF (the second scenario), and minority of the matched TFs have a cobinding with the target TF (the first scenario), which demonstrates that FCNA is inclined to first find the TFs belonging to the same TF family and then the cobinding TFs.

Refining the prediction performance by using the located regions

In this section, we explored how to utilize the located regions efficiently. Firstly, the trained FCNA was used to locate the regions of 50 bp that may contain TFBSs. Secondly, five threshold values {0.5, 0.6, 0.7, 0.8 and 0.9} were set to transform the probabilities of outputs into 1 or 0, and the locating accuracy was computed in terms of the Equation (4). Thirdly, according to the locating accuracy ($A_{0.5} = 0.871$, $A_{0.6} = 0.896$, $A_{0.7} = 0.907$, $A_{0.8} = 0.921$ and $A_{0.9} = 0.938$), the value 0.9 was adopted to filter out a few false-positive regions from the located regions, and the rest of

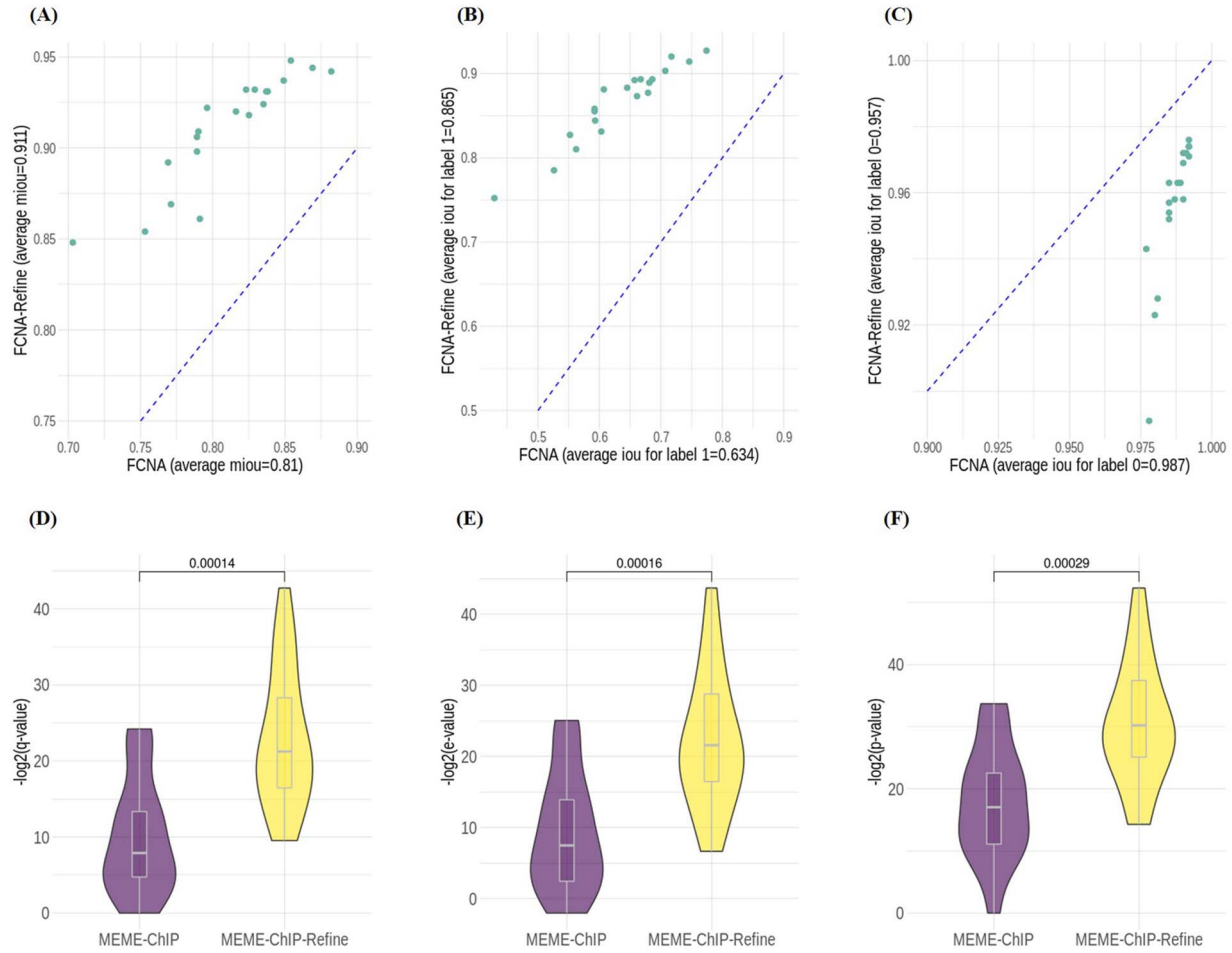


Figure 5. The performance comparison of methods and refined methods, including FCNA and FCNA-Refine (A, B, C), MEME-ChIP and MEME-ChIP-Refine (D, E, F).

them was used as the final located regions. Finally, MEME-ChIP and a slightly-modified FCNA that just adjusts the kernel size of convolutional layers and pooling layers were employed to predict motifs on the final located regions. In the experiments, the 20 TF-DNA binding datasets from A549 were used, and the detailed results were recorded in Supplementary Table 6.

$$A = \frac{\sum \left(seq_i = \begin{cases} 1, TP \cap PP \text{ is true} \\ 0, \text{otherwise} \end{cases} \right)}{N}, i \in [1, \dots, N] \quad (4)$$

where TP and PP separately represent the true positive sites (label 1) and the predicted positive sites of the i -th sequence, and N denotes the total number of sequences.

As shown in Figure 5 A-C, we used the slightly-modified FCNA to conduct experiments on the located regions, named as FCNA-Refine, and tested the miou, the iou for label 0 (0-iou) and the iou for label 1 (1-iou). From the results, the miou of FCNA-Refine is much better than that of FCNA, and the miou gain is 0.101. The 1-iou of FCNA-Refine is much better than that of FCNA, and the gain is 0.231, whereas 0-iou of FCNA-Refine is worse than that of FCNA, and the gain is -0.03, of which the reason is that the number of label 0 in the located regions is

much less than that in the original sequences. However, we were more concerned about the iou for label 1.

As shown in Figure 5 D-F, we used MEME-ChIP to conduct experiments on the located regions, named as MEME-ChIP-Refine, and tested the $-\log_2(P\text{-value})$, $-\log_2(e\text{-value})$ and $-\log_2(q\text{-value})$. From the results, the three statistical significances of MEME-ChIP-Refine are significantly higher than the values for MEME-ChIP (Wilcoxon test P-values are 0.00014, 0.00016 and 0.00029, respectively).

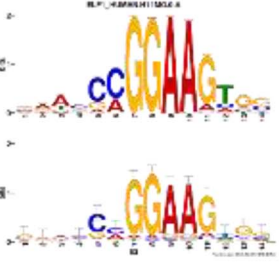
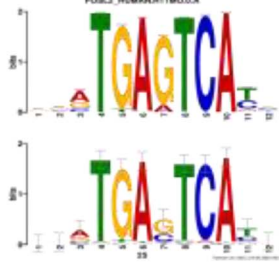
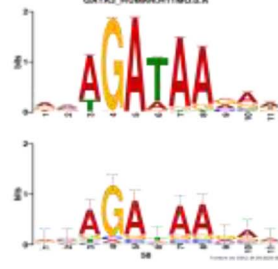
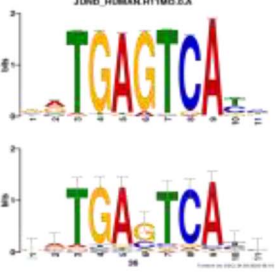
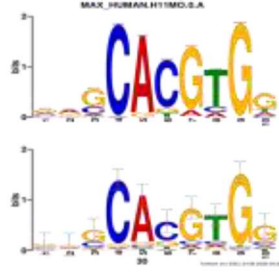
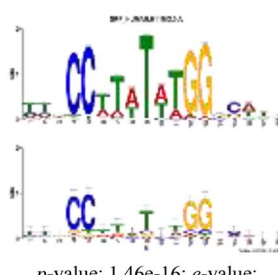
Above all, we find that the regions located by FCNA can be used by motif discovery tools to further refine the prediction performance. Thereby, FCNA can be employed to locate target regions from coarse data, and the located regions are then used for the subsequent analysis.

Identifying TF-DNA binding motifs across different cells

In this section, we investigated the performance of FCNA for identifying TF-DNA binding motifs across different cell lines. In the experiments, FCNA was trained on the A549 and GM12878 cells to predict motifs on the MCF7 cell.

As shown in Table 2 and Supplementary Table 2, the P-value, e-value, q-value, and miou are very high, and the predicted motif logos are also well matched with the experimentally validated

Table 2. Motif logos visualization

<p>ELF1</p>  <p>p-value: $9.80\text{e-}15$; e-value: $3.93\text{e-}12$; q-value: $7.67\text{e-}12$; miou: 0.838</p>	<p>FOSL2</p>  <p>p-value: $6.85\text{e-}16$; e-value: $2.75\text{e-}13$; q-value: $5.29\text{e-}13$; miou: 0.906</p>	<p>GATA3</p>  <p>p-value: $1.19\text{e-}10$; e-value: $4.77\text{e-}08$; q-value: $9.24\text{e-}08$; miou: 0.779</p>
<p>JUND</p>  <p>p-value: $7.12\text{e-}13$; e-value: $2.85\text{e-}10$; q-value: $5.49\text{e-}10$; miou: 0.851</p>	<p>MAX</p>  <p>p-value: $4.87\text{e-}11$; e-value: $1.95\text{e-}08$; q-value: $3.72\text{e-}08$; miou: 0.852</p>	<p>SRF</p>  <p>p-value: $1.46\text{e-}16$; e-value: $5.85\text{e-}14$; q-value: $1.15\text{e-}13$; miou: 0.816</p>

motifs in the HOCOMOCO database, which demonstrates that FCNA can accurately identify TF-DNA binding motifs across different cell lines. Thereby, FCNA can be used to predict the same TF-DNA binding motifs across different cell lines.

Conclusions and discussion

In this paper, we proposed a novel motif discovery method, namely FCNA which incorporates a FCN, a global average pooling, and a hard negative mining loss, to locate TFBSs and predict TF-DNA binding motifs. Experimental results on several ChIP-seq datasets show that FCNA significantly outperforms the competing methods, which demonstrates that FCNA can efficiently solve the problem of failing to accurately predict motifs that almost all DL-based methods face. Besides, through a series of experiments, we find that (i) FCNA is inclined to first find the TFs belonging to the same TF family and then find the cobinding TFs; (ii) the regions located by FCNA can be used by motif discovery tools to further refine the prediction performance; (iii) FCNA can accurately identify TF-DNA binding motifs across different cell lines and thereby be used to predict the same TF-DNA binding motifs across different cell lines.

Two possible issues in this paper should be discussed: (i) since it is difficult to use FCNA to precisely locate TFBSs, so the results predicted by FCNA must contain a few false-positive samples. To remove these false-positive samples, a high threshold value was used in the experiments, but doing this will inevitably delete some true positive samples; (ii) since FCNA makes use of the strongly-supervised label information (nucleotide-level

labels) to predict motifs, so FCNA is overwhelmingly dependent on the quality of nucleotide-level labels. Therefore, some more comprehensive methods should be proposed to solve the two issues in the future works.

Key Points

- In the task of predicting transcription factor binding sites (TFBSs), most of methods based on deep learning mainly focus on predicting the sequence specificity of TF-DNA binding and fail to identify motifs and TFBSs accurately. It is important to develop an efficient method for solving this problem.
- The concept of using fully convolutional networks to locate TFBSs and predict binding motifs is firstly proposed in this paper.
- Experimental results on *in vivo* datasets show that the proposed method significantly outperforms other competing methods and that the located regions can be used to further improve the prediction performance.

Acknowledgment

This work was supported by the grant of National Key R&D Program of China (No. 2018AAA0100100 & 2018YFA0902600) and partly supported by National Natural Science Foundation of China (Grant nos. 61861146002, 61732012, 62002266, 61772370, 61932008, 61772357 and 62073231) and supported

by 'BAGUI Scholar' Program and the Scientific & Technological Base and Talent Special Program, GuiKe AD18126015 of the Guangxi Zhuang Autonomous Region of China and supported by Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), LCNBI and ZJLab.

References

1. Vaquerizas JM, Kummerfeld SK, Teichmann SA, et al. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;10:252–63.
2. Elnitski L, Jin VX, Farnham PJ, et al. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res* 2006;16:1455–64.
3. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* 2012;13:840–52.
4. Bailey TL, Williams N, Misleh C, et al. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 2006;34:W369–73.
5. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;27:1653–9.
6. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011;27:1696–7.
7. 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.
8. Zhang Q, Zhu L, Bao W, et al. Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. *IEEE/ACM Trans Comput Biol Bioinform* 2020;17:679–89.
9. Zhang Q, Yu W, Han K, et al. Multi-scale capsule network for predicting DNA-protein binding sites. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Institute of Electrical and Electronics Engineers, 2020.
10. 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.
11. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12:931–4.
12. 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.
13. Kulakovskiy IV, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* 2018;46:D252–9.
14. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proc IEEE Conf Comput Vision Pattern Recognit* 2015;3431–40.
15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv 1409.1556* 2014.
16. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vision Pattern Recognit* 2016;770–8.
17. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Munich Germany: Springer, 2015, 234–41.
18. Yu C, Wang J, Peng C, et al. Learning a discriminative feature network for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City Utah USA: Proc IEEE Conf Comput Vision Pattern Recognit, 2018, 1857–66.
19. Yu C, Wang J, Peng C, et al. Bisenet: bilateral segmentation network for real-time semantic segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*, Munich Germany: Springer, 2018, 325–41.
20. Mathelier A, Xin B, Chiu T-P, et al. DNA shape features improve transcription factor binding site predictions in vivo. *Cell systems* 2016;3:278–286. e274.
21. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, Montreal, Canada: MIT Press, 2015, 91–9.
22. Gupta S, Stamatoyannopoulos JA, Bailey TL, et al. Quantifying similarity between motifs. *Genome Biol* 2007;8:R24.
23. Zhang Q, Shen Z, Huang D-S. Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. *Sci Rep* 2019;9:1–12.
24. Zhang Q, Shen Z, Huang D-S. Predicting in-vitro transcription factor binding sites using DNA sequence+ shape. *IEEE/ACM Trans Comput Biol Bioinform* 2019.
25. Wang J, Zhuang J, Iyer S, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012;22:1798–812.
26. Isakova A, Groux R, Imbeault M, et al. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat Methods* 2017;14:316.