Molecular Therapy
Nucleic Acids
Original Article

Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture

Siguo Wang,[1] Qinhu Zhang,[1,2] Zhen Shen,[3] Ying He,[1] Zhen-Heng Chen,[4] Jianqiang Li,[4] and De-Shuang Huang[1]

[1]The Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, No. 4800 Caoan Road, Shanghai 201804, China; [2]Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University, Siping Road 1239, Shanghai 200092, China; [3]School of Computer and Software, Nanyang Institute of Technology, Changjiang Road 80, Nanyang, Henan 473004, China; [4]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

The study of transcriptional regulation is still difficult yet fundamental in molecular biology research. Recent research has shown that the double helix structure of nucleotides plays an important role in improving the accuracy and interpretability of transcription factor binding sites (TFBSs). Although several computational methods have been designed to take both DNA sequence and DNA shape features into consideration simultaneously, how to design an efficient model is still an intractable topic. In this paper, we proposed a hybrid convolutional recurrent neural network (CNN/RNN) architecture, CRPTS, to predict TFBSs by combining DNA sequence and DNA shape features. The novelty of our proposed method relies on three critical aspects: (1) the application of a shared hybrid CNN and RNN has the ability to efficiently extract features from large-scale genomic sequences obtained by high-throughput technology; (2) the common patterns were found from DNA sequences and their corresponding DNA shape features; (3) our proposed CRPTS can capture local structural information of DNA sequences without completely relying on DNA shape data. A series of comprehensive experiments on 66 *in vitro* datasets derived from universal protein binding microarrays (uPBMs) shows that our proposed method CRPTS obviously outperforms the state-of-the-art methods.

INTRODUCTION

Protein-DNA interactions play an important role in the regulation of gene transcription, splicing, translation, and degradation.[1–3] The binding of transcription factors (TFs) and DNA is the basic molecular mechanism in gene regulation. TF binding sites (TFBSs) on DNA are short sequences located in regulatory regions of genes, typically range from a few to about 20 base pairs (bp), and binding regions of one TF on different genes are usually conservative. When given an input DNA sequence, classifying whether or not there is a binding site for a particular TF is a core task of bioinformatics. The identification of TFBSs, also known as motif discovery (MD) problems, is usually defined as finding similar subsequences from a given set of DNA sequences. Unfortunately, only some of these binding sites have been identified by expensive and time-consuming biological experiments. With the rapid development of high-throughput sequencing technology, numerous *in vitro* experimental data were provided by protein binding microarrays (PBMs), which is very important, as we stand by the reliability data, and provides the possibility to improve the computational method for TF binding specific expression and binding site discovery.[4–8] Compared with biological experiment methods, computational methods have become the main method for solving burning biological questions[9–11] due to their advantages of simplicity, speed, and low cost.

Due to the intricate motif variations, indirect effects on binding specificity, and noisy data, it is difficult to infer motifs from high-throughput data.[12–15] Position weight matrix (PWM) is the most earliest and common method to characterize the specificity of protein-DNA binding sequences due to its simplicity and comprehension,[16] which allows an intuitive visualization as a sequence logo.[17] However, it still has certain limitations, considering the dependence between nucleotides, kmer-SVM and gapped k-mer, has been proposed based on k-mer features successively.[18,19] Gapped k-mer was combined with deep neural networks to solve the limitations of kernel skills in kmer-SVM. Besides, many studies have shown that sequence specificity can be better captured using more complex models.[20] Since deep learning (DL) can automatically find predictive signatures and process the input high-dimensional data,[21,22] the technology has reached rapid development and demonstrated extraordinary performance in various tasks, including but not limited to speech recognition,[23] computer vision,[24] natural language processing,[25] and functional genomics prediction.[26,27] DL has achieved unprecedented

Received 19 November 2020; accepted 14 February 2021;
https://doi.org/10.1016/j.omtn.2021.02.014.

Correspondence: De-Shuang Huang, The Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, No. 4800 Caoan Road, Shanghai 201804, China.
E-mail: dshuang@tongji.edu.cn
Correspondence: Qinhu Zhang, The Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, No. 4800 Caoan Road, Shanghai 201804, China.
E-mail: 20310031@tongji.edu.cn

performance for capturing motif patterns and elucidating complex regulatory mechanisms based on large-scale chromatin immunoprecipitation sequencing (ChIP-seq) datasets.[28,29] A variety of computational approaches have been developed for predicting TFBSs from the raw DNA sequences based on DL. A couple of pioneering studies have come up with the idea that a convolutional neural network (CNN) can be applied to genomics relying on the basic building blocks of CNN in computer vision.[24] Deepbind is a hallmark of the first successful application of a model sequence of CNNs to predict the specificity of protein binding.[28] There are also a few hybrid DL models used in DNA sequences, of which DanQ is a typical example to quantify the function of DNA sequences.[30] Furthermore, a variety of computational approaches has been developed based on novel convolutional architectures; for example, circular filters were proposed to efficiently utilize data and easily interpret learned filters.[31] A systematic evaluation of DL architectures for predicting DNA- and RNA-binding specificity, named deepRAM, provided in-depth analysis for the researcher.[32]

Although these DL methods have made great achievements, there are still some drawbacks, ignoring that DNA is a complex three-dimensional macromolecule. As a result of the advances in DNA structure elucidation, four distinct DNA shape features, including minor groove width (MGW), propeller twist (ProT), helix twist (HelT), and roll, can be computationally derived from DNA sequences by Monte Carlo (MC) simulation.[33] To explain the intricacies of the DNA structure, 13 features were expanded in subsequent research, which includes six intra-bp parameters, six inter-bp parameters, and one MGW.[34] These data provide unprecedented opportunities to predict TFBS and capture more features that affect TF binding. Recent research shows that adding DNA shapes plays a significant role in simulating and predicting TF-DNA binding affinities, which adopted traditional machine learning methods.[35] Yang et al.[36] used a simple neural network framework combined with DNA shape to predict motifs from human ChIP-seq data. It is worth noting that a sequence + shape framework (DLBSS), based on DL, was proposed, which used a shared CNN to find common patterns from DNA sequences and their corresponding DNA shape features.[37] However, despite the DL models or the traditional methods perform well by adding DNA shape to identify the TFBSs, the ability to come up with effective methods to improve the performance of capturing sequence-specific motif and predict specific TF binding to genomic DNA remains a challenge.

With the consideration that although DLBSS applied the CNN model based on the combination of DNA shape and DNA sequences, which has achieved certain achievements, there are still some defects that need to be improved, such as data bias, extracting local dependencies between motifs, visualizing binding motifs of TFs, etc. Hence, we introduce a shared hybrid DL architecture inspired by Zhang et al.,[37] a strategy of combining CNN and recurrent neural network (RNN), adapting the DNA sequences and their corresponding local DNA shape features in this paper. CNN was used to capture low-level spatial information from given DNA sequences and DNA shape features, and RNN was used to capture long-term dependencies between sequences. The CRPTS framework is mainly composed of data input, model training, and binding intensity output. The architecture diagram of CRPTS is shown in Figure 1. Our comprehensive experiments on 66 *in vitro* universal PBM (uPBM)[14,35] datasets demonstrate that CRPTS significantly outperforms some existing state-of-the-art methods in the prediction of TF binding affinity. Last but not least, CRPTS has the ability to discover some essential experimentally verified binding motifs, and the key point of our method provides verified binding motifs to visualize and interpret our models. More details regarding the datasets, the model architecture, and the evaluation metrics used in our work can be found in the following sections.

RESULTS

In this section, we describe the evaluation performance of the proposed model compared with the state-of-the-art approaches. We carried out a series of experiments on 66 *in vitro* datasets, and the experimental results show that CRPTS significantly performs better than the competing methods.

Performance evaluation and comparison

To prevent overfitting and experimental accuracy, the performance of the CRPTS was measured using 5-fold cross-validation for each of 66 *in vitro* datasets. In order to evaluate the proposed method, we compared the performance of predicting TFBSs with recent state-of-the-art methods. The quality of the model for predicting binding affinity was evaluated by using the coefficient of determination (R^2) and Pearson correlation coefficient (PCC), which were applied in Weirauch et al.[14] We assumed that the closer the two evaluation indexes are to 1, the better the method is. We used the two indicators for each dataset, and we calculated the average of the 66 datasets about two indicators to verify the comprehensive performance level of the method. Two performance measures are defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - Y_i)^2}{\sum_i (y_i - \overline{y})^2} \quad \text{(Equation 1)}$$

$$PCC(y, Y) = \sum_i \frac{y_i - \overline{y}}{\sqrt{(y_i - \overline{y})^2}} \cdot \frac{Y_i - \overline{Y}}{\sqrt{\left(Y_i - \overline{Y}\right)^2}}, \quad \text{(Equation 2)}$$

where $y_i, Y_i, \overline{y}$, and $\overline{Y}$, respectively, represent the observed, predicted, average observed, and average predicted binding affinity scores.

Accuracy comparison and analysis

In order to evaluate the performance of CRPTS more synthetically, we not only compare CRPTS with Deepbind, which only used the primary DNA sequences as the inputs based on the CNN model,[28] but also with three methods that combined DNA sequences and DNA shapes, including two kernel-based methods (spectrum + shape kernel, di-mismatch + shape kernel)[35] and a DL method-based DLBSS.[37] We compared the performance of CRPTS with the competing methods on 66 *in vitro* datasets regarding the metrics R^2

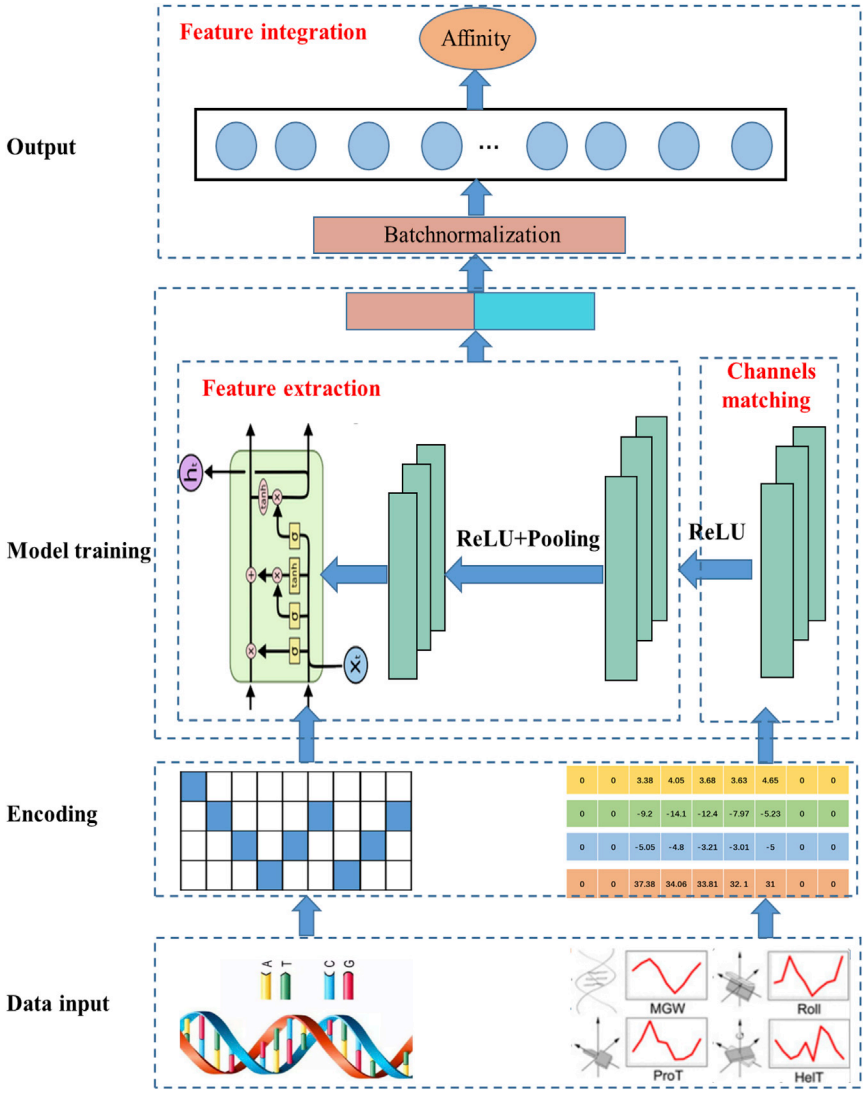


Figure 1. The general architecture of the CRPTS

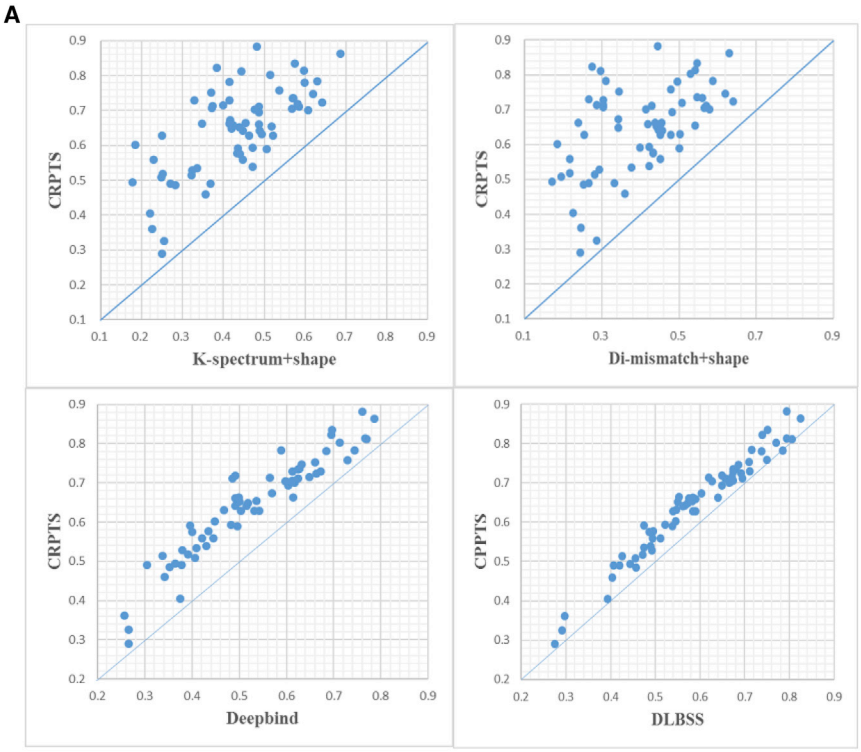
(1) Data input. The input data includ DNA sequence data and four shape features. (2) Encoding. DNA sequences were converted into a matrix by one-hot encoding, and DNA shape features were processed by a sliding window to obtain the input matrix. (3) Model training. The matrix of DNA shape features was first input into the convolutional layer to match the numbers of channels; then two types of data were input to the shared hybrid model to extract features. (4) Output. The features were combined by two fully connected layers after a batch normalization layer to obtain the final affinity.

and PCC mentioned above. A detailed comparison of the two evaluation criteria R^2 and PCC for CRPTS and competing methods is shown in Tables S1 and S2.

Furthermore, Figure 2 plots an overall performance comparison of CRPTS and the competing methods on 66 *in vitro* datasets. As shown in Figure 2, it is evident that CRPTS achieves a more remarkable and stable performance than the competing methods in terms of PCC and R^2. A couple of observations are notable from these plots: CRPTS is clearly better than the two kernel-based methods, which demonstrates that the DL model combining DNA sequence and DNA shape information has a significant effect on identifying TFBSs. As indicated in Figure 3, CRPTS achieves a statistically significant improvement in average R^2 and PCC. CRPTS achieves about 6% and 4% higher than DLBSS with respect to R^2 and PCC, which shows that our proposed hybrid DL model has a clear advantage over the one using only CNN. It is worth noting that both maximum and minimum values of CRPTS are better than that of the competing method in 66 *in vitro* datasets. Generally speaking, the smaller box of the proposed method CRPTS indicates that the range of the two indicators is more concentrated, which proves that the proposed method has strong stability. The reasons for the outstanding performance of CRPTS may lie in that (1) CRPTS explicitly considers the shape information of DNA sequences, and (2) CRPTS uses long short-term memory (LSTM) to further extract long-term dependencies between DNA shape features and DNA sequences.

The effect of DNA shape on the identification of TFBSs

In order to examine whether adding DNA shape information to the shared hybrid model can improve the prediction accuracy of TF binding affinities, we extended an experiment only using independent DNA sequences as input based on the shared hybrid model, named CRPT. Then we analyzed the binding event prediction capability of CRPTS (using DNA sequence + DNA shape features as data input) and CRPT (only using DNA sequence as data input) to exhibit the superiority of our model. It can be seen from Figure 4, the median of CRPTS is slightly higher than CRPT, and the data distribution is consistent, which indicates that the stability of the model is not affected by the data input. As shown in Figure 3, DLBSS is far superior to Deepbind (the average values of R^2 and PCC has increased by 6% and 3%, respectively). The reason for the high performance gain lies in the following: (1) Deepbind only consists of one convolutional layer that is used to score all potential motifs, but the local structural information of DNA sequences is not considered, and (2) DNA shape features are explicitly combined in the DLBSS for considering the local structural information. Compared to CRPT, the performance of CRPTS is slightly improved (the average values of R^2 and PCC has increased by 1.3% and 1%); the details are shown in the Tables S1 and S2. The reason for low performance gain lies in the following: CRPT consists of a convolutional layer and a



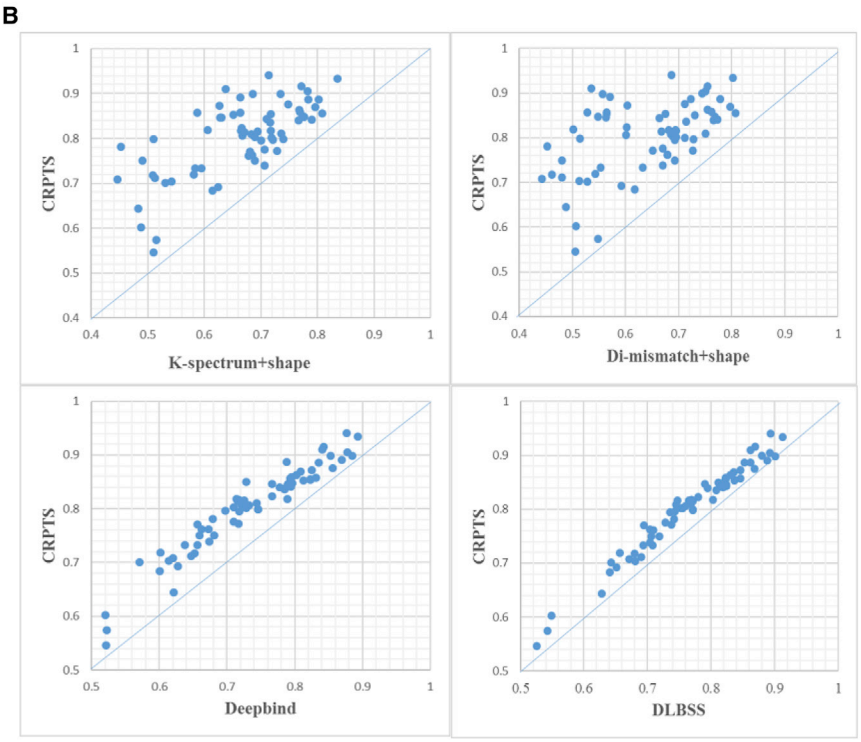


Figure 2. An overall performance comparison about R^2 and PCC of CRPTS and the competing methods on 66 *in vitro* datasets

(A) An overall performance comparison about R^2. (B) An overall performance comparison about PCC.

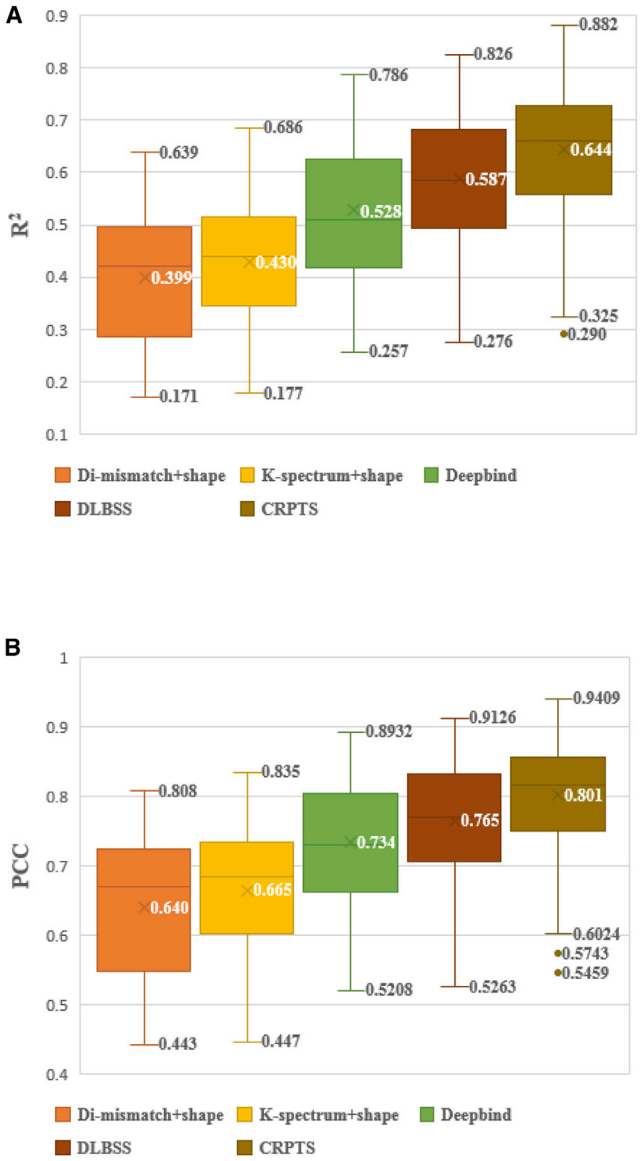


Figure 3. Boxplots for R^2 and PCC value of CRPTS and the competing methods

(A) Boxplots for R^2. (B) Boxplots for PCC.

LSTM layer, where the convolutional layer is used to score all potential motifs, and the LSTM layer is used to learn the local structural information and long-term dependence in the sequences, rather than completely relying on DNA shape data. Compared with Deepbind and DLBSS, it is no surprise that CRPTS and CRPT can capture most of the structural information from the raw sequence data without completely relying on DNA shape data.

Identifying and visualizing binding motifs of TFs

Model visualization is crucial in computational biology. Several methods have been proposed to explain the parameters of neural networks and gain insights into the learned characteristics. CNN not only can efficiently process the DNA sequence but also can automatically extract features, and the kernels are similar to PWMs, describing the popular model TF of sequence-specific binding. In order to further evaluate this model, we visualized the identified binding motifs of TFs. For each kernel, the sequences of the position having the highest activation value that is greater than zero are selected and collected and are used to make a file in multiple expectation maximizations for motif elicitation (MEME) motif format. The similarities between identified motifs and true motifs are calculated by Gupta et al.[38] and is publicly available at the MEME Suite of motif analysis tools.[39,40] All kernels in the convolutional layer are initialized randomly, and all motifs are automatically learned during model training. The experimental results show that CRPTS achieves higher accuracy on *in vitro* datasets compared to the competing methods. As shown in Table 1, several motifs learned by CRPTS and the corresponding motifs in the standard database are recorded.

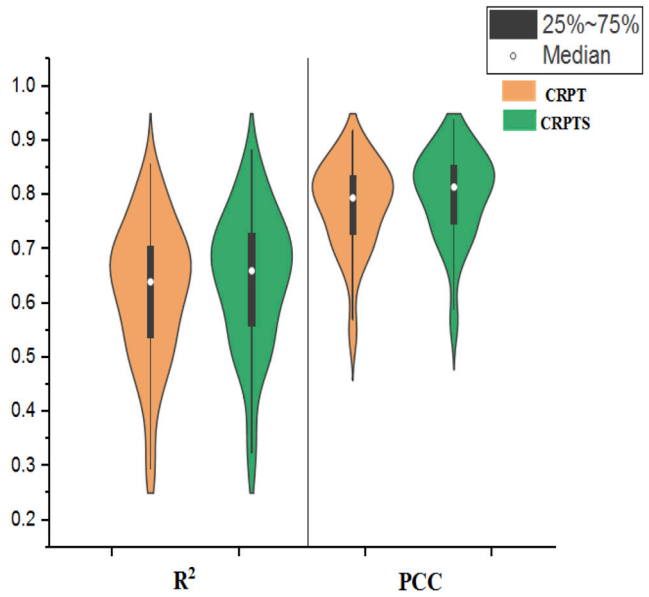
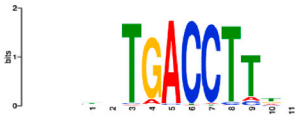
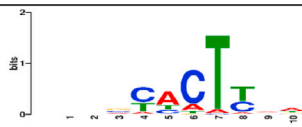
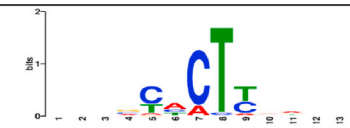


Figure 4. The effect of DNA shape on the R^2 and PCC

DISCUSSION

Although there are many computational methods to identify TFBSs, it still remains a major challenge for the research community. Recent research suggests that DNA shape features play an important role in the recognition of DNA binding sites,[41] to date, but a systematic and comprehensive method is lacking. Therefore, a shared hybrid CNN/RNN architecture CRPTS combining DNA sequences and DNA shapes was proposed to identify TFBSs. To verify the performance of CRPTS, a series of experiments on 66 *in vitro* uPBM datasets were performed in this paper, and experimental results demonstrated that the proposed method effectively improves the prediction performance of TFBSs and outperformed some state-of-the-art methods in terms of R^2 and PCC. Moreover, the comparison

Table 1. Examples of TF motifs learned by the hybrid CRPTS models and compared with the motifs recorded in UniPROBE

UniPROBE	CRPTS	Information
		p-value:4.50e-06 E-value:4.34e-03 q-value:4.25-03 name:Dbp
		p-value:1.06-06 E-value:1.02e-03 q-value:3.39e-04 name:Foxo6
		p-value:3.05e-06 E-value:1.81e-03 q-value:3.62e-03 name:Oct1
		p-value:1.58e-03 E-value:1.53+00 q-value:5.50e-01 name:Nr2f2
		p-value:4.49e-11 E-value:2.67e-08 q-value:1.67e-08 name:Nkx2-9

between the learned motifs and the standard motifs in the Universal PBM Resource for Oligonucleotide Binding Evaluation (UniPROBE) database shows a high degree of matching.

The main contributions of our proposed CRPTS lie in the following: (1) a shared hybrid model combining DNA sequences and DNA shape features was first used in the identification of TFBSs, CNN was used to identify spatial information from given DNA sequences and shape characteristics, and RNN mainly captured long-term dependencies between sequence motifs; (2) the common patterns were found from DNA sequences and their corresponding DNA shape features, which were then concatenated to compute a predicted affinity value; and (3) experiment results show that CRPTS can extract local structural information rather than completely relying on extra DNA shapes.

This design of the framework is flexible in accommodating any type of high-throughput data, including but not limited to the data discussed in this paper. There are several questions worthy pursuing following our work, trying to combine 13 DNA structure features with DNA sequences to improve the performance of TFBSs. Nine additional DNA shape features were derived from three different data sources, including MC, molecular dynamics, and X-ray crystallography (XRC) experiments.[42] Similar to the four shapes, the pentamer characteristic value of the additional shape was also obtained, and we will make attempts to combine open chromatin ATAC data and ChIP-seq data based on a shared hybrid model to provide some theoretical basis for genomic research. It is worth noting that some experimental results have shown that the basic polar and hydrophobic lle amino acids of principal properties play an important role in modulating the interaction between TFs and methylated DNA,[43,44] which provides a new insight for combining the physicochemical properties of DNA and DL to improve the interpretability of TFs. The ultimate goal of our research will be used to discover disease targets and provide a basis for accurate diagnosis and treatment of complex diseases.[45–47]

MATERIALS AND METHODS

Datasets and data preprocessing

66. uPBM data

The PBM as a rapid, high-throughput sequence technology not only provides a biochemical representation of TF-DNA interactions

in vitro but also provides biological insights into the *in vivo* functions and regulatory roles of TFs. The PBM technology has enabled a genome scale to characterize the sequence specificities of DNA-protein interactions in a high-throughput manner. In order to evaluate the performance of the proposed CRPTS, we downloaded 66 uPBM data from the Dialogue for Reverse Engineering Assessments and Methods 5 (DREAM5),[14] which comes from a variety of protein families. Each TF dataset consists of more than 40,000 arrays containing all patterns of unaligned 35-mer probes and a complete set of PBM probe intensities from two distinct microarray designs, named HK and ME. In order to be suitable for a DL model, each input DNA sequence was converted into a matrix $n \cdot l$ by one-hot encoding, where n corresponds to the four nucleotides A, T, C, and G, represented by binary vectors A = [1,0,0,0], C = [0,0,1,0], T = [0,1,0,0], and G = [0,0,0,1], respectively, and l equals 35 in uPBM that we used.

DNA shape data

The three-dimensional structure of DNA plays an important role in determining the DNA binding preferences of TFs[48,49] and other DNA-binding proteins.[50,51] In previous work, the unique pentamers of four DNA shape features include the following: MGW, roll, ProT, and HelT, obtained from MC simulations by a sliding-window approach together with a query table.[33] The original DNA shape data was provided in Table S3. A pentamer contributes one MGW value, one ProT value, two roll values, and two HelT values, and we take the average of the two roll and HelT values, one MGW value, and ProT as the final effective value of the four DNA shapes. In order to make the input type of the DNA sequence and corresponding DNA shape feature consistent, we padded two zeros at both sides of the sequence to make the length 1 + 4, and then the sliding window was used to obtain the shape feature matrix $n \cdot l$, where n represents the number of shapes, and l represents the length of the sequence. For the sake of eliminating the bias that was caused by different ranges of values for different shapes, we normalized each feature by zero-mean normalization, respectively, as follows:

$$x' = \frac{x - \mu}{\sigma}, \qquad \text{(Equation 3)}$$

where x is the original feature value of pentamer, and x' is the normalized value. μ and σ are the mean and standard deviation of all samples, respectively.

Model construction and training

The network architecture of the proposed CRPTS

Overall, we proposed a novel shared hybrid CNN and RNN framework to improve the accuracy of predicting TFBSs. The hybrid model in CRPTS consists of channels matching module, feature extraction module, and feature integration module.

Channels matching module. This module is composed of a convolutional layer and a ReLU (rectified linear unit) layer. Since DNA sequences and DNA shape are heterogeneous data, it is necessary to apply a reasonable strategy to integrate them. In this paper, we have used four shapes, but they were actually expanded to 13 repertoires recently.[34] In order to integrate more shape information into our model, we first applied the same strategy as DLBSS that used a convolutional layer with kernel size 1 to process the DNA shape features to match the number of channels of the hybrid network. More specifically, our model requires binary as input, as described in Datasets and data preprocessing; a DNA sequence was converted into an image-like matrix $n \cdot l$ using one-hot encoding, where n corresponds to the four nucleotides A, G, T, and C, and l represents the length of DNA sequence. For each convolutional network layer, the outputs of the layer were calculated by the following formula:

$$conv(X)_{ik} = f\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{mn}^{k} X_{i+m,n} + b_{ik}\right), \qquad \text{(Equation 4)}$$

where X is the input, i is the index of the output position, and k is the index of kernels. W is the convolutional weight tensor and can be interpreted as a four shape motif detector; b is the bias term. The convolution outputs were transformed by the activation function $f(\cdot)$,which is an element-wise nonlinear function ReLU in our model. ReLU is an activation function widely used in DL, which can alleviate gradient vanishing problems during back-propagation training and has better convergence performance. ReLU is defined as follows:

$$ReLU(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}. \qquad \text{(Equation 5)}$$

A more detailed design of channels matching module is shown in Table 2.

Table 2. The detailed settings of channels matching module

Architectures	Settings	Output shape
Convolutional layer	kernel number = 4, kernel size = 1, stride = 1, padding = 0	(B, n, 4)
ReLU layer	–	(B, n, 4)

Feature extraction module. This model is composed of a convolutional layer, a ReLU layer, a max pooling layer, an LSTM layer, and a dropout layer. The DNA shape information processed by CNN and original DNA sequence data were input into the shared hybrid neural network to extract features for predicting TFBSs. The advantage of applying a shared model is that it not only greatly reduces the parameters of the network but also can be trained in parallel. First, the convolution layer followed by a ReLU layer with kernel 16 captured spatial information from given DNA sequences and DNA shape features. One of the advantages of applying CNNs to analyze omics data is multiple types of input data, features can be easily integrated, and the effective features can be automatically discovered by representation learning. Then, a max pooling layer was used to pick out the maximum value based on the outputs from the convolutional

Table 3. The detailed settings of feature extraction module

Architectures	Settings	Output shape
Convolutional layer	kernel number = 16, kernel size = 13, stride = 1, padding = 6	(B, n, 16)
ReLU layer	–	(B, n, 16)
Max-pooling layer	global	(B, 16)
LSTM layer	32	(B, 32)
Dropout layer	0.2	(B, 32)

layer, which reduces the dimension of the input to make the model computationally efficient. The pooling operation is defined as:

$$pooling(X)_{ik} = \max\left(\left\{X_{iM,k}, X_{(iM+1),k}, \ldots, X_{(iM+M-1),k}\right\}\right). \quad \text{(Equation 6)}$$

The max pooling layer was followed by an LSTM layer to capture long-term dependencies between the motifs and the orientations and spatial distances between sequences. As we have learned, RNN is an alternative to CNNs for processing sequential data, and LSTM networks are first proposed to use special hidden units to remember inputs for long periods. The key of LSTM is the cell state, which is carefully regulated by structures called gates, including an input gate, a forget gate, and an output gate. In the first step, the "forget gate" decides what information is to be discarded or saved. The next step is to decide how much new information should be added to the cell state. The final step determines what value to output.

$$f_t = sigmoid\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad \text{(Equation 7)}$$

$$i_t = sigmoid(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Equation 8)}$$

$$C_t = tanh(W_G \cdot [h_{t-1}, x_t] + b_G) \quad \text{(Equation 9)}$$

$$S_t = f_t \odot S_{t-1} + i_t \odot C_t \quad \text{(Equation 10)}$$

$$O_t = sigmoid(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Equation 11)}$$

$$h_t = O_t \odot tanh(S_t), \quad \text{(Equation 12)}$$

where W represents the weight matrix; b represents the bias; f_t, i_t, and O_t represent the weight values of the forget, input, and output gates; x_t, C_t, and h_t represent the input vector, the memory representation, and the hidden layer state at time t, respectively; and $\odot$ is element-wise multiplication. A dropout[52] layer with a probability of 0.2 was added to avoid overfitting by ignoring half of the feature detectors and to make the model have a stronger generalization ability. Next, all of the dropout results of both DNA sequences and DNA shape information were combined into a feature vector, which is then input to the output stage. A more detailed design of feature extraction module is shown in Table 3.

Table 4. The detailed settings of feature integration module

Architectures	Settings	Output shape
Concatenation layer	–	(B, 32)
Batchnormalization layer	–	(B, 32)
Fully connected layer	unit number = 32	(B, 32)
ReLU layer	–	(B, 32)
Dropout layer	ratio = sample from {0.2, 0.5, 0.7}	(B, 32)
Fully connected layer	unit number = 1	(B, 1)

Feature integration module. This module is composed of two fully connected layers, a batch normalization layer, and a dropout layer. In the output stage, batch normalization was applied before inputting it into the fully connected layer, which not only avoids a gradient problem during backpropagation but also simplifies the initialization process of the parameters of the network.[53] The outputs from the batch normalization were then fed into a fully connected layer with 32 hidden neurons to integrate features. The output layer followed by a dropout layer containing only one neuron was used to predict the TF-DNA binding specificity. A more detailed design of feature integration module is shown in Table 4.

The network architecture of the proposed CRPTS

For each dataset, we minimize the reasonable loss function of mean squared error (MSE) to train the proposed hybrid model. The loss function is defined as follows:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\overline{y_i} - y_i^2\right) + \lambda\|\theta\|_2, \quad \text{(Equation 13)}$$

where $\overline{y_i}$ and y_i, respectively, indicate the ground and estimated signal intensity a, and N is the number of sequences in each training dataset. L2 regularization was used to avoid model overfitting, λ means a regularization parameter, and $\|\bullet\|_2$ means L2 norm. We use AdaDelta to optimize the loss function and set the mini-batch size as 300. The dropout ratio, momentum, and Delta in AdaDelta in the neural network were randomly selected from [0.2,0.5], [0.9,0.99,0.999], and [1e–8,1e–6,1e–4], respectively. To prevent overfitting and ensure experimental accuracy, we used five-fold cross-validation. The optimal parameter set was retained during the training process and applied to the entire training dataset, and the epochs of training were set at 100. Moreover, an early stopping strategy was adopted to reduce the running time. The source code is freely provided in the GitHub repository (https://github.com/wangguoguoa/CRPTS).

Our implementation is written in Python, utilizing the Keras 2.0 library with the TensorFlow 1.2.0[54] backend. We used a Linux machine with 32 gigabytes (GBs) of memory and an NVIDIA Titan X Pascal graphics processing unit (GPU) for training.

Generating motifs learned by the proposed model

Various DL algorithms have brought about breakthroughs in solving specific types of problems in genomic applications that extract

important features from the raw data, but one of the main drawbacks is their poor interpretability. In the neural network, a convolution kernel is similar to the motif detector, taking both DNA sequences and structure information into account, the activation position of the kernel is located by scanning all of the sequences, and each kernel is converted into a PWM. For each sequence in each dataset, we found a position that had the maximum convolution value among all the kernels, and a 13-bp (the size of the kernel) subsequence starting at this position was extracted. As described in Model construction and training, the number of convolution kernels is 16, and the size is 13 in our proposed CRPTS. Therefore, the frequencies of the four nucleotides at each position were then calculated, and the 16 PWM matrix representing the TF motif was derived.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2021.02.014.

ACKNOWLEDGMENTS

This work was supported by grants from the National Key R&D Program of China (2018AAA0100100 and 2018YFA0902600), partly supported by grants from the National Natural Science Foundation of China (61861146002, 61732012, 61772370, 61932008, 61772357, 62002266, 62073231, and 62002297), supported by the "BAGUI Scholar" Program and the Scientific & Technological Base and Talent Special Program, GuiKe AD18126015 of the Guangxi Zhuang Autonomous Region of China, the Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), LCNBI, and ZJLab.

AUTHOR CONTRIBUTIONS

S.W., Q.Z., J.L., and D.-S.H. conceived the study and carried out analyses. Z.S. and Y.H. collected the data samples and information. S.W., Q.Z., and Z.-H.C. wrote and revised the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Huang, D.S., Zhang, L., Han, K., Deng, S., Yang, K., and Zhang, H. (2014). Prediction of protein-protein interactions based on protein-protein correlation using least squares regression. Curr. Protein Pept. Sci. *15*, 553–560.
2. Zhu, L., Deng, S.P., and Huang, D.S. (2015). A Two-Stage Geometric Method for Pruning Unreliable Links in Protein-Protein Networks. IEEE Trans. Nanobioscience *14*, 528–534.
3. Zhu, L., You, Z.-H., Huang, D.-S., and Wang, B. (2013). t-LSE: a novel robust geometric approach for modeling protein-protein interaction networks. PLoS ONE *8*, e58368.
4. Guo, J., Lofgren, S., and Farrel, A. (2014). Structure-based prediction of transcription factor binding sites. Tsinghua Sci. Technol. *19*, 568–577.
5. Huang, D.-S., and Yu, H.-J. (2013). Normalized Feature Vectors: A Novel Alignment-Free Sequence Comparison Method Based on the Numbers of Adjacent Amino Acids. IEEE/ACM Trans. Comput. Biol. Bioinform. *10*, 457–467.
6. Huang, D.-S., Zhao, X.-M., Huang, G.-B., and Cheung, Y.-M. (2006). Classifying protein sequences using hydropathy blocks. Pattern Recognit. *39*, 2293–2300.
7. Deng, S.-P., and Huang, D.-S. (2014). SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method. Methods *69*, 207–212.
8. Xia, J.-F., Zhao, X.-M., Song, J., and Huang, D.-S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics *11*, 174.
9. Huang, D.S., and Zheng, C.H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics *22*, 1855–1862.
10. Zheng, C.-H., Zhang, L., Ng, V.T., Shiu, S.C., and Huang, D.-S. (2011). Molecular pattern discovery based on penalized matrix decomposition. IEEE/ACM Trans. Comput. Biol. Bioinformatics *8*, 1592–1603.
11. Huang, D.-S., and Jiang, W. (2012). A General CPL-AdS Methodology for Fixing Dynamic Parameters in Dual Environments. IEEE Trans. Syst. Man Cybern. B Cybern. *42*, 1489–1500.
12. Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. *4*, 393–411.
13. Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol. *12*, 918–922.
14. Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al.; DREAM5 Consortium (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.
15. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.
16. Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23.
17. Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.
18. Fletez-Brant, C., Lee, D., McCallion, A.S., and Beer, M.A. (2013). kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res. *41*, W544–W556.
19. Cao, Z., and Zhang, S. (2020). Probe Efficient Feature Representation of Gapped K-mer Frequency Vectors from Sequences Using Deep Neural Networks. IEEE/ACM Trans. Comput. Biol. Bioinformatics *17*, 657–667.
20. Kazan, H., Ray, D., Chan, E.T., Hughes, T.R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput. Biol. *6*, e1000832.
21. Deshuang, H., and Songde, M. (1996). A new radial basis probabilistic neural network model. Proceedings of Third International Conference on Signal Processing, *2*, pp. 1449–1452.
22. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.
23. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhouke, V., Nguyen, P., Sainath, T.N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process. Mag. *29*, 82–97.
24. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Commun. ACM *60*, 84–90.
25. He, X., Gao, J., and Deng, L.; Microsoft Research (2014). Deep learning for natural language processing and related applications (Tutorial at ICASSP), https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/ICASSP_DeepTextLearning_v07.pdf.
26. Park, Y., and Kellis, M. (2015). Deep learning for regulatory genomics. Nat. Biotechnol. *33*, 825–826.
27. Huang, D.-S., and Du, J.-X. (2008). A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans. Neural Netw. *19*, 2099–2115.
28. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

29. Zhu, L., Guo, W.-L., Deng, S.-P., and Huang, D.-S. (2016). ChIP-PIT: Enhancing the Analysis of ChIP-Seq Data Using Convex-Relaxed Pair-Wise Interaction Tensor Decomposition. IEEE/ACM Trans. Comput. Biol. Bioinformatics *13*, 55–63.

30. Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. *44*, e107.

31. Blum, C.F., and Kollmann, M. (2019). Neural networks with circular filters enable data efficient inference of sequence motifs. Bioinformatics *35*, 3937–3943.

32. Trabelsi, A., Chaabane, M., and Ben-Hur, A. (2019). Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics *35*, i269–i277.

33. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.

34. Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res. *45*, 12877–12887.

35. Ma, W., Yang, L., Rohs, R., and Noble, W.S. (2017). DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. Bioinformatics *33*, 3003–3010.

36. Yang, J., Ma, A., Hoppe, A.D., Wang, C., Li, Y., Zhang, C., Wang, Y., Liu, B., and Ma, Q. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res. *47*, 7809–7824.

37. Zhang, Q., Shen, Z., and Huang, D.-S. (2019). Predicting in-vitro transcription factor binding sites using DNA sequence + shape. IEEE/ACM Trans. Comput. Biol. Bioinform. Published online October 15, 2019. https://doi.org/10.1109/TCBB.2019.2947461.

38. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

39. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

40. Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E. (1997). Meta-MEME: motif-based hidden Markov models of protein families. Comput. Appl. Biosci. *13*, 397–406.

41. Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

42. Chiu, T.-P., Xin, B., Markarian, N., Wang, Y., and Rohs, R. (2020). TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. *48* (D1), D246–D255.

43. Shen, Z., and Zou, Q. (2020). Basic polar and hydrophobic properties are the main characteristics that affect the binding of transcription factors to methylation sites. Bioinformatics *36*, 4263–4268.

44. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

45. Deng, S.-P., Zhu, L., and Huang, D.-S. (2015). Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks. BMC Genomics *16* (*Suppl 3*), S4.

46. Zheng, C.H., Huang, D.S., Zhang, L., and Kong, X.Z. (2009). Tumor clustering using nonnegative matrix factorization with gene selection. IEEE Trans. Inf. Technol. Biomed. *13*, 599–607.

47. Deng, S.P., Zhu, L., and Huang, D.S. (2016). Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks. IEEE/ACM Trans. Comput. Biol. Bioinformatics *13*, 27–35.

48. Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530–543.

49. Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. *3*, 1093–1104.

50. West, S.M., Rohs, R., Mann, R.S., and Honig, B. (2010). Electrostatic interactions between arginines and the minor groove in the nucleosome. J. Biomol. Struct. Dyn. *27*, 861–866.

51. Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. *24*, 814–826.

52. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

53. Ioffe, S., and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine, *37*, pp. 448–456.

54. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, pp. 265–283.