Construction and Prediction of Antimicrobial Peptide Predicition Model Based on BERT

Jia-Qi Zong[1], Zheng-Qiu Yu[1], Yu-Qing Gong[1], Min-Shu Wang[1], Pei-Xuan Lin[1]

[1]24520211154668, [1]24520210157079, [1]24520211154654
[1]24520211154663, [1]23020201153948

Abstract

Predicting enhancer-promoter interactions (EPIs) task, is of great significance for understanding gene regulation and grasping the mechanism of a disease. In recent years, computational methods based on machine learning have been widely used in predicting EPIs because of their good performance. Although existing models have achieved some achievements, there are still some problems. For example, a model cannot learn more sequence information; the structure of a model is simple, and it is difficult to extract more effective features. To solve the above problems, We propose a novel method, termed EPI-DLMH, for predicting EPIs with the use of DNA sequences only. EPI-DLMH consists of three major steps. First, a two-layer convolutional neural network is used to learn local features, and a bidirectional gated recurrent unit network is used to capture long-range dependencies on the sequences of promoters and enhancers. Second, an attention mechanism is used for focusing on relatively important features. Finally, a matching heuristic mechanism is introduced to explore the interaction between enhancers and promoters. We use benchmark datasets in evaluating and comparing the proposed method with existing methods. Comparative results show that our model is superior to currently existing models in all cell lines.

Introduction

Research in recent years has shown that non-coding DNA can not be regarded as "junk", but perform a variety of important biological functions, such as assisting gene regulation and providing signal functions (Esteller 2011). Among them, enhancers are one of the important non-coding elements, and they play a central role in the regulation of gene expression (Shlyueva, Stampfel, and Stark 2014). In the past few decades, the identification of EPIs has mainly relied on high-throughput experimental techniques, but these traditional experimental methods have limitations such as low resolution and low throughput, and the complex mechanism of action of EPIs makes traditional methods very time-consuming Laborious.

With the update and iteration of computer technology and the advent of the era of big data, deep learning technology has been proven to be an effective data mining technology. Deep learning can automatically learn and extract important features from a large amount of labeled data, and finally discover the laws that exist behind important features. Nowadays, various high-throughput technologies have produced a large amount of biological sequence data, which can provide enough training samples for deep learning technology. Deep learning has been widely used in various research fields in the biological field (Miotto et al. 2017), such as cancer detection based on gene expression data (Fakoor et al. 2013), protein-ligand interaction prediction (Wang et al. 2014), protein secondary structure prediction (Spencer, Eickholt, and Cheng 2015), Human non-coding RNA identification (Fan and Zhang 2015), etc.

With the increase of biological data and the development of computer technology, more and more researchers choose to use biological computing methods to replace traditional biological experimental methods. Many computational methods have been used to quickly and accurately identify EPIs in multiple cells lines. Deep learning methods based on DNA sequences are one of the important methods. The deep learning method based on DNA sequence mainly uses deep learning technology to extract features of the DNA sequence information of enhancers and promoters, and use sequence features to predict the EPIs present in the genome. The biggest advantage of this method is that researchers can predict EPIs efficiently and accurately using only limited DNA sequence information. Mao et al. (Mao, Kostka, and Chikina 2017) proposed a deep learning model EPIANN based on the attention mechanism. EPIANN can only use DNA sequences to predict EPIs and achieve good performance.

Related work

Studies have shown that chromatin's three-dimensional (3D) organization plays an important role in transcriptional regulation (Zhang et al. 2019). Among them, the identification of EPIs is an effective way to understand gene regulation and grasp disease mechanisms (Li et al. 2012). Therefore, EPIs is an important step towards a deeper understanding of gene regulation and disease mechanisms. Many computational methods have been used to rapidly and accurately identify EPIs in multiple cell lines, and we focus on deep learning methods based on DNA sequences. Mao et al.(Mao, Kostka, and Chikina 2017) proposed a deep learning model EPI-

ANN based on the attention mechanism. Mao et al. (Mao, Kostka, and Chikina 2017) proposed EPIANN, a deep learning model based on attention mechanism, which can predict EPIs using only DNA sequences and achieve good performance. The SPEID model not only has good prediction performance but also has a fast training speed. Later, Zhuang et al. (Zhong, Xiaotong, and Wei) constructed a prediction model SIMCNN using only CNN structure by simplifying the SPEID model, which has a prediction performance close to that of SPEID but has a great improvement in training speed, and most importantly, also applies migration learning to the study of predicting EPIs. In conclusion, deep learning methods based on DNA sequences have yielded good results in predicting EPIs.

However, these models have the following shortcomings:first, all models use one-hot coding as feature embedding, which has some defects (such as prone to dimensional disaster, ignoring the relationship between subsequences, etc.); econd, the existing model architecture is very simple, it is difficult to learn effective features. Third, in all models, the feature representation of a pair of sequences is directly connected for subsequent prediction, so it is easy to lose the potential interactive information between the two; fourth, the existing deep learning model has many parameters, resulting in low training efficiency of the model.

Proposed solution

Structure of EPI-DLMH

We propose a model EPI-DLMH based on deep learning and heuristic matching. The model can only use DNA sequences to predict EPIs. The EPI-DLMH model structure is shown in Figure 1. The four main steps in the model are sequence embedding, feature extraction, heuristic matching, and prediction. To put it simply, for a pair of enhancer sequence and promoter sequence as input, they are divided and expressed as a feature matrix, and then different multi-layer hybrid neural networks are used to extract more feature information. Then the two feature vectors are connected using different heuristic matching strategies and finally put into a prediction layer to predict EPIs. This process can determine whether there is an interaction between the enhancer and the promoter, and the details of the model will be introduced later.

Sequence embedding

The k-mer representation is an efficient method for analyzing long DNA sequences. In the experiments in this chapter, a sliding window of k bp length was set up according to the k-mer representation, a sliding window of k bp length was set, and s was used as the sliding step to split the enhancer sequences and promoter sequences. Since it is mentioned in the literature (Yang et al. 2017), PEP-WORD verified in the experiments that the computational efficiency of the model and the information complexity of the vector are optimal when k=6.

Therefore, in the experiments, we set k=6 and s=1. With the k-mer representation, the sequence AGCTGTTC can be split into AGCTGT, GCTGTT, and CTGTTC correspondingly. The k-mer representation is easy to understand and compute. All EPIs prediction models use one-hot encoding to embed the subsequence features after splitting. This method directly encodes the subsequence as a one-hot vector, this feature representation can easily cause dimensional disasters.

EPI-DLMH represents k-mers words using the Dna2vec (Ng 2017) model. Dna2vec is based on the popular language model Word2vec, which is trained on a shallow two-layer neural network. Dna2vec can obtain low-dimensional and high-quality vectors to represent k-mers words. In the experiments in this chapter, dna2vec is first pre-trained with chr1 to chr22 chromosomal genes on the hg38 human body and then fine-tuned to fit the EPIs prediction task using the EPIs dataset. In this way, Dna2vec can obtain a 100-dimensional vector corresponding to all 6-mers words. Finally, the model can encode the input promoter and enhancer sequences into feature matrices of size 100 * 2000 and 100 * 3000 (since the enhancer and promoter sequence lengths are fixed to 3000 bp and 2000 bp, respectively).

Feature Extraction

In the EPI-DLMH model, a hybrid network structure containing CNN (Ren et al. 2019) (Li and Liu 2019) and Bi-GRU is used for feature extraction. CNN is mainly used to learn the local features of the promoter and enhancer sequences, while Bi-GRU is mainly used to capture the long-term dependence of the local features. In addition, an attention layer is added to compute the important features and assign larger weights to the more important features, which are finally represented as feature vectors. The model first constructs a two-layer CNN consisting of a convolutional layer and a max pooling layer. The convolutional layer is mainly used to learn the local features of enhancers and promoters, and the max pooling layer is used for dimensionality reduction. In the experimental setting of this chapter, different convolutional and pooling layers are built for enhancers and promoters, respectively. For the enhancer, the filter length of the convolutional layer, the number of filters, the pooling length of the max pooling layer, and the pace are set to 60, 64, 30, 30. And for the initiator, they are set to 40, 64, 20, 20, respectively. The hyperparameter settings are consistent with the comparison models (SPEID and SIMCNN).

The local features output from the above CNN are then fed into Bi-GRU, which is a bidirectional recurrent neural network that captures bidirectional semantic dependencies. Here, the Bi-GRU layer is used to learn the long-term dependencies of local features. Specifically, the Bi-GRU layer has two parts that simultaneously learn features from the forward and reverse directions. The update process for GRU is as follows:

$$r_t = \sigma\left(W_r X_t + U_r h_{t-1} + b_r\right) \tag{1}$$

$$z_t = \sigma\left(W_z X_t + U_z h_{t-1} + b_z\right) \tag{2}$$

$$\bar{h}_t = \tanh\left(W_h x_t + U_h\left(r_t \cdot h_{t-1}\right) + b_h\right) \tag{3}$$

$$h_t = (1 - z_t) \cdot \bar{h}_t + z_t \cdot h_{t-1} \tag{4}$$

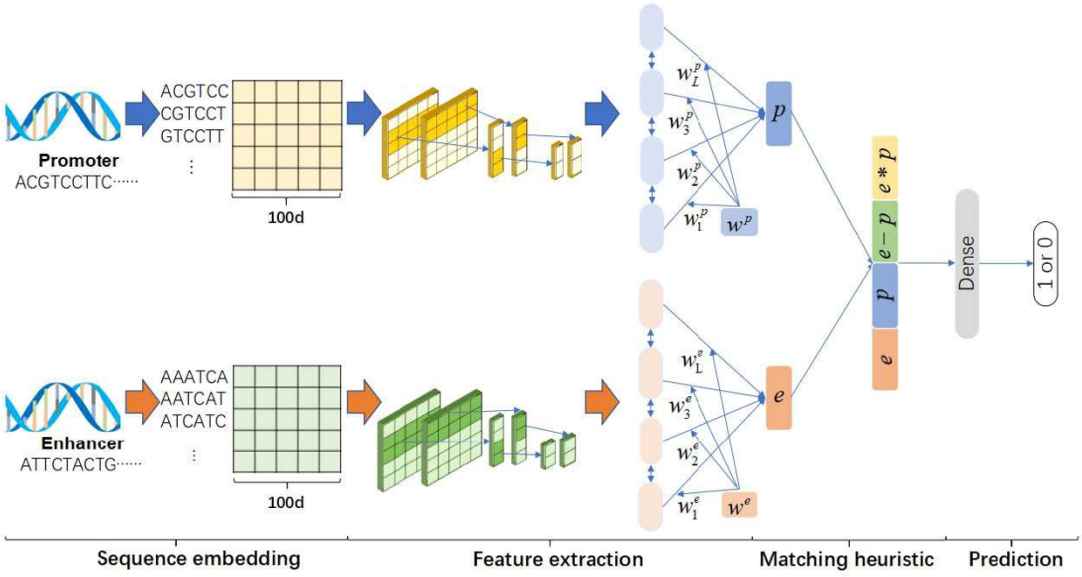

Figure 1: Framework of our EPI-DLMH model

$$h_t = \left[\vec{h}_t, h_t\right] \tag{5}$$

where $tanh$ is the hyperbolic tangent function and σ is the sigmoid function, U is the weight matrix of the previous hidden layer vector h_{t-1}, x_t is the input at time t, $\bar{h}_t$ is the candidate activation probability, $\cdot$ is the product of two elements, r_t and z_t respectively Indicates reset gate vector and update gate vector. Bi-GRU generates the forward feature sequence $\left\{\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_t, ..., \overrightarrow{h}_L\right\}$ from left to right and generates the reverse feature sequence $\left\{\overleftarrow{h}_1, \overleftarrow{h}_2, ..., \overleftarrow{h}_t, ..., \overleftarrow{h}_L\right\}$ from right to left. Finally, the forward feature and the reverse feature are combined. In the experimental environment of this chapter, different Bi-GRUs are used to train enhancers and promoters, and both are set to 50 output units. The Attention Mechanism originated from the study of vision in humans. In the EPI-DLMH model, the attention layer is used to learn the importance of features, so that more weights are assigned to the more important features. The specific formula is as follows:

$$w_k^p = \frac{\exp\left(\rho_{kp}^T b_\varepsilon\right)}{\sum_{k=1}^{L} \exp\left(\rho_{kp}^T b_\varepsilon\right)} \tag{6}$$

$$w_k^e = \frac{\exp\left(\rho_{ke}^T b_\varepsilon\right)}{\sum_{k=1}^{L} \exp\left(\rho_{ke}^T b_\varepsilon\right)} \tag{7}$$

$$p = \sum_{k=1}^{L} w_k^p * h_k^p \tag{8}$$

$$e = \sum_{k=1}^{L} w_k^e * h_k^e \tag{9}$$

where ρ_{kp}^T and ρ_{ke}^T are the hidden layer representations of the k-th feature of the promoter and enhancer, respectively. b_ε is the context vector, w_k^p and w_k^e represent the importance of h_k^p and h_k^e, h_k^p and h_k^e are the output of the promoter and enhancer in the Bi-GRU layer at time k, and e and p represent the feature vector of the enhancer and the promoter. In the experimental environment, different attention layers are used for enhancers and promoters, and both are set to 50 output units.

Heuristic matching

To capture more interactive information between promoters and enhancers, EPI-DLMH introduces an additional heuristic matching mechanism, which has been widely used in various advanced models of natural language. Some research compared the combination of different sentence encoders and heuristic matching mechanisms in the Natural Language Inference (NLI) task. The sentence pairs in the NLI task are first expressed as feature vectors by the sentence encoder. Then the two are combined and put into the prediction layer using a heuristic matching algorithm. The experimental results show that heuristic matching can improve task performance. As for the EPIs prediction task, it can also be regarded as a sentence pair task in natural language processing, such as machine question answering, sentence similarity, natural language reasoning, and so on.

So in EPI-DLMH, the model uses heuristic matching to capture the inner relationship between promoters and enhancers. Specifically, EPI-DLMH introduces three different heuristic matching strategies: (1) the connection of enhancer features and promoter features; (2) the difference between enhancer features and promoter features; (3) Product of enhancer features and promoter features. The first type of heuristic matching is used to connect the features of enhancers and promoters, the second type of heuristic matching is used to calculate the proximity of features between en-

hancers and promoters, and the third type of heuristic matching is used to calculate the similarity of characteristics between enhancers and promoters. After passing the feature extraction stage, the model will obtain the feature vector corresponding to the enhancer and the feature vector corresponding to the promoter, and then calculate and connect the three heuristic matches between them. The specific formula is as follows:

$$m = [e, p, e - p, eop] \tag{10}$$

where e and p represent the feature vector of enhancer and promoter respectively, $-$ is the element difference, o is the product of the elements and m is the product of the elements.

Experiment

To compare the EPI-DLMH model with other existing models (SPEID, SIMCNN, EPIANN), the same datasets are used in the experiment and the same training process is followed to train all models. For any given cell line, the training process is described as follows:

1. Start from the unbalance dataset D.
2. Split D into training set D_{train}(90% of D) and test set D_{test}(10% of D) by Stratified sampling.
3. Enhance D_{train} to get a balance dataset D_{aug}.
4. Train the model on D_{aug}.
5. Test the model on D_{test} and evaluate.

EPI-DLMH uses the Glorot algorithm (Glorot and Bengio 2010) to initialize the weights of each network in the experiments in this chapter. The model incorporates Dropout and L2 regularization in some layers to prevent overfitting. The model is trained on 32 small-batch samples by setting the cross-entropy loss function and minimizing the loss using the Adam algorithm (Kingma and Ba 2014). The model is trained in 90 epochs in each cell line, in line with the comparison models (SPEID, SIMCNN, EPIANN).

Data acquisition and preprocessing

The benchmark data set contains six different cell lines, namely GM12878, HeLa-S3, IMR90, K562, HUVEC, and NHEK. There is a serious data imbalance in each cell line. The ratio of positive samples to negative samples is about 1:20, as shown in Table 1. Unbalanced data will affect the judgment of traditional classifiers and make its prediction performance biased. To solve this problem, we use upsampling to amplify the number of positive samples using the upstream and downstream regions of the enhancer (positive) and the upstream and downstream regions of the promoter (positive). The length of the expansion window is three kbps and two kbps, respectively, so that the expansion window is sufficient to include all enhancer sequences and related regions around the promoter sequence.

We first evaluate the impact of the introduced heuristic matching mechanism on the model performance, and different combinations of heuristic matching are used in the experiments to evaluate and compare different EPI-DLMH models; secondly, we compare the performance of EPI-DLMH models with the existing models SPEID, SIMCNN and EPIANN; finally, we explore the intrinsic mechanism of performance improvement and the future improvement direction of EPI-DLMH models.

Cell Line	Positive Sample	Negative Sample
GM12878	2113	42200
HUVEC	1524	30400
HeLa-S3	1740	34800
IMR90	1254	25000
K562	1977	39500
NHEK	1291	25600

Table 1: Detailed distribution of the six cell line datasets

	GM12878	HOVEC	Hela-S3	IMR90	K562	NH3K
SPEID	0.937	0.855	0.847	0.879	0.937	0.969
SIMCNN	0.937	0.924	0.941	0.946	0.951	0.967
EPIANN	0.937	0.924	0.918	0.945	0.943	0.959
EPI-DLMH(best)	**0.949**	**0.948**	**0.952**	**0.948**	**0.955**	**0.977**

Table 2: AUROC values of different models in six cell lines

	GM12878	HOVEC	Hela-S3	IMR90	K562	NH3K
SPEID	0.796	0.606	0.733	0.753	0.780	0.889
SIMCNN	0.786	0.642	0.791	0.733	0.805	0.887
EPIANN	0.723	0.702	0.616	0.770	0.673	0.861
EPI-DLMH(best)	**0.819**	**0.720**	**0.824**	**0.818**	**0.826**	**0.893**

Table 3: AUPRC values of different models in six cell lines

	GM12878	HOVEC	Hela-S3	IMR90	K562	NH3K
SPEID	0.764	0.535	0.718	0.731	0.692	0.734
SIMCNN	0.726	0.541	0.652	0.709	0.732	0.850
EPIANN	0.699	0.639	0.590	0.711	0.626	0.797
EPI-DLMH(best)	**0.766**	**0.649**	**0.780**	**0.778**	**0.795**	**0.861**

Table 4: F1 values of different models in six cell lines

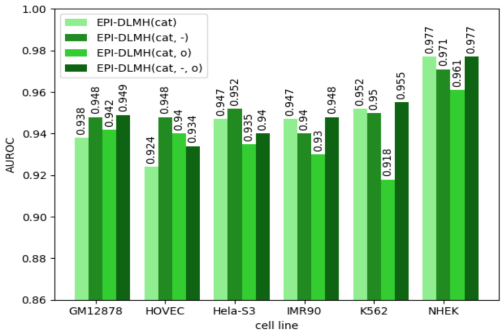

Figure 2: AUROC values for EPI-DLMH models

Heuristic matching improves predictive performance

This section shows whether additional heuristic matching mechanisms can improve model performance. Different EPI-DLMH models were evaluated and compared using different heuristic matching combinations in the experiment. Figures 2, 3, and 4 show the scores of AUROC, AUPRC, and F1 based on different combinations of heuristic matches, where "Cat" refers to the Association of enhancers with the characteristics of promoters; "-" and "o" represent the differences and product of the characteristics of the enhancer and the promoter, respectively. As can be seen from the graph, the additional heuristic matching algorithm can effectively

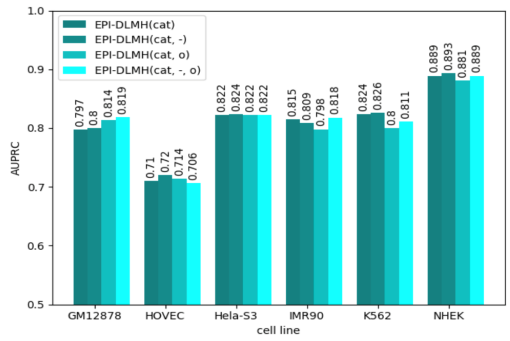

Figure 3: AUPRC values for EPI-DLMH models

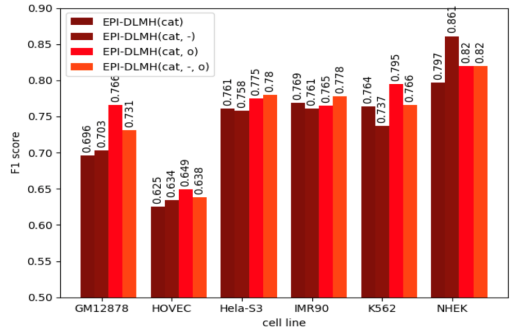

Figure 4: F1 values for EPI-DLMH models

improve the prediction performance in most cases (multi-cell lines). For example, in Nhek Cell Line, the F 1 index of EPI-DLMH (cat, -) was 6.4% higher than that of EPI-DLMH (cat). However, in other cell lines, the model's performance was not improved by introducing heuristic matching. For example, in the K562 cell line, the AUPRC of EPI-DLMH (cat,-, O) was lower than that of EPI-DLMH (cat), and the F1 index was lower than that of EPI-DLMH (cat). According to the experiments in this chapter, the (cat, -) or (cat,-, O) combinations in heuristic matching perform best in most cases. And for a particular cell line, different heuristic matches often perform differently and can be superior to direct-link training. Therefore, the experimental results show that the proposed heuristic matching algorithm can effectively capture the potential interaction between enhancer and promoter, thus improving the model performance.

Performance comparison

To evaluate the effectiveness of the EPI-DLMH model, this section compares it with the three most advanced EPIs prediction models, namely SPEID, SIMCNN, and EPIANN, using a benchmark data set of six cell lines. This section illustrates the model's performance on the three evaluation indicators of AUROC, AUPRC, and F1 index in table 2, table 3, and table 4, respectively.

Table 2 shows the performance comparison of the EPI-DLMH model with other existing models in terms of AUROC metrics. The performance of EPI-DLMH in each cell line was better than that of other models (SPEID, SIMCNN, and EPIANN). Table 3 shows the performance comparison of the EPI-DLMH model with other models in terms of AUPRC metrics.Combining tables 4-2 and 4-3, it was found that although the EPI-DLMH model showed only a slight increase in AUROC in IMR90 and K562 fine cell lines , and the increase in AUPRC was larger in these two cell lines. Table 4 shows the performance of the EPI-DLMH model and other models on F1 indicators. Specifically, the EPI-DLMH model has a certain performance improvement compared with the second model, 0.2% on GM12878, 1% on Hevec, 6.2% on Helamuri S3, 4.7% on IMR90, 6.3% on K562, 1.1% on NHEK.

In summary, EPI-DLMH showed better performance than any other model in each cell line. Because EPI-DLMH constructs an efficient network structure, the sequence features of web-based learning are more abundant than other models, and can accurately capture the structural information between the enhancer sequence and the promoter sequence. And the heuristic matching mechanism introduces more interactive information between enhancer and promoter to further improve the model's prediction performance. In addition, this section also compares the training time of the EPI-DLMH model with the existing model. This chapter uses the same 1080Ti GPU to train all the models. Among them, the training time of the contrast model is 180h(SPEID), 120h(SIMCNN), and 240h(EPIANN), which is much longer than that of EPI-DLMH(72h). The reason is that EPI-DLMH uses fewer parameters when building a network model. To sum up, the EPI-DLMH model has better prediction performance and higher training efficiency than other existing models.

Conclusion

This paper proposes a deep learning and heuristic matching-based model EPI-DLMH, which can predict EPIs using only DNA sequences. The EPI-DLMH model is divided into sequence embedding, feature extraction, heuristic matching, and prediction steps. The experimental results show that EPI-DLMH has better prediction performance and faster training in multiple cell lines than existing models. And heuristic matching was proved to improve the prediction performance of the model. For EPI-DLMH learns richer sequence features than other models, and the introduced heuristic matching mechanism can bring more interactions of enhancer and promoter sequences mutual information. However, EPI-DLMH still needs certain improvements in future work, such as the heuristic matching mechanism. Although it has achieved excellent results in various fields, the interpretability of the mechanism itself is insufficient and still needs more theoretical and experimental argumentation. The pre-training process of EPI-DLMH is based on the Word2vec model. With the development of NLP technology, more and more language models are emerging, such as ELMO, XLNet, etc. The study shows that XLNet is more suitable for long sequences of corpus than language models such as BERT and Word2vec, which helps to improve the subsequent task performance. Therefore, newer pre-trained language models can be tried to improve the prediction results.

References

Esteller, M. 2011. Non-coding RNAS in human diseases. *Nature Reviews Genetics* 12(12): 861–874.

Fakoor, R.; Ladhak, F.; Nazi, A.; and Huber, M. 2013. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning*, volume 28, 3937–3949. ACM, New York, USA.

Fan, X. N.; and Zhang, S. W. 2015. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular Biosystems* 11(3): 892–897.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* .

Li, C. C.; and Liu, B. 2019. MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Briefings in Bioinformatics* .

Li, G.; Ruan, X.; Auerbach, R.; Sandhu, K.; Zheng, M.; Ping, W.; Poh, H.; Goh, Y.; Lim, J.; and Zhang, J. 2012. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* 148(1-2): 84–98.

Mao, W.; Kostka, D.; and Chikina, M. 2017. Modeling enhancer-promoter interactions with attention-based neural networks. *bioRxiv* 219667.

Miotto, R.; Fei, W.; Shuang, W.; Jiang, X.; and Dudley, J. T. 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19(6).

Ng, P. 2017. dna2vec: Consistent vector representations of variable-length k-mers .

Ren, F.; Yang, C.; Qiu, Q.; Zeng, N.; and Zou, Q. 2019. Exploiting Discriminative Regions of Brain Slices Based on 2D CNNs for Alzheimer's Disease Classification. *IEEE Access* 7(99): 181423–181433.

Shlyueva, D.; Stampfel, G.; and Stark, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15(4): 272.

Spencer, M.; Eickholt, J.; and Cheng, J. 2015. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *Computational Biology Bioinformatics IEEE/ACM Transactions on* 12(99): 103–112.

Wang, C.; Liu, J.; Luo, F.; Tan, Y.; Deng, Z.; and Hu, Q. N. 2014. Pairwise input neural network for target-ligand interaction prediction. *Yamaguchi Journal of Economics Business Administrations Laws* 41.

Yang, Y.; Zhang, R.; Singh, S.; and Ma, J. 2017. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics* .

Zhang, W.; Li, W.; Zhang, J.; and Wang, N. 2019. Data Integration of Hybrid Microarray and Single Cell Expression Data to Enhance Gene Network Inference. *Current Bioinformatics* 14(3): 255–268.

Zhong, Z.; Xiaotong, S.; and Wei, P. ???? A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. *Bioinformatics* (17): 2899–2906.