Prediction of enhancer-promoter interactions using the cross-cell information and domain adversarial neural network

Fang Jing[1], Shao-Wu Zhang[1*] and Shihua Zhang[2,3,4*]

[1]Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[2]NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

* Corresponding authors. Emails: zhangsw@nwpu.edu.cn; zsh@amss.ac.cn

Abstract

Background: Enhancer-promoter interactions (EPIs) play key roles in transcriptional regulation and disease progression. Although several computational methods have been developed to predict such interactions, their performances are not satisfactory when training and testing data from different cell lines. Currently, it is still unclear what extent EPI prediction across cell lines can be made based on sequence-level information.

Results: In this work, we present a novel Sequence-based method (called SEPT) to predict the Enhancer-Promoter interactions in the new cell line by using the cross-cell information and Transfer learning. SEPT first leans the features of enhancer and promoter from DNA sequences with convolutional neural network (CNN), then designing the gradient reversal layer of transfer leaning to reduce the cell line specific features meanwhile retaining the features associated with EPIs. When the locations of enhancers and promoters are provided in new cell line, SEPT can successfully recognize EPIs in this new cell line based on labeled data of other cell lines. The experiment results show that SEPT can effectively learn the latent import EPIs-related features between cell lines and achieves the best prediction performance in terms of AUC (the area under the receiver operating curves).

Conclusion: SEPT is an effective method for predicting EPIs in the new cell line. Domain adversarial architecture of transfer leaning used in SEPT can learn the latent EPI features shared among cell lines from all other existing labeled data. It can be expected that SEPT will be of interest to researchers concerned with biological interaction prediction.

Keywords: Enhancer-promoter interactions, Cell line, Convolutional neural network, Transfer learning, Gradient reversal layer

1 Background

The enhancer-promoter interactions (EPIs) play a critical role in gene regulation in eukaryotes. In genetics, a promoter is a region of DNA sequence upstream of a particular gene [1]. The length of a promoter is probably hundreds to thousands of base pairs [2]. Its function aims to initiate gene transcription of a particular gene. While a enhancer is also an important transcriptional regulatory short DNA fragments that further activate the level of transcription of its target genes by contacting close physical proximity to the promoters in the three-dimensional (3D) nuclear space [3]. Hundreds of thousands of enhancers have been estimated to be contained in the human genome. Normally, a promoter is under the control of multiple enhancers, and multiple promoters can be regulated by a single enhancer. Additionally, the distances between interacting enhancer and promoter pairs have varied widely, varying from kilobases to millions of base-pairs because of the chromatin folding in the 3D space [4-8]. Moreover, as more and more studies reported that enhancer sequence variations are associated with serious human diseases [9-12]. Thus, the importance of EPIs for gene expression is matter-of-course.

Over the past decade, many high-throughput experimental approaches, such as chromosome conformation capture-based (3C) [13] and its variants of Hi-C [14] and ChIA- PET [1], have been developed to study the chromatin interactions. Although Hi-C and ChIA-PET could measure the whole genome DNA-DNA interactions, the

genomic resolutions are often low, varying from few kilobases to tens of thousands bases [15-17]. In order to study EPIs, very high (<10kb) resolution data is needed. All these experimental approaches are technically challenging, time-consuming and have high false-negative rate. What is more, the EPIs vary across different cellular conditions and tissues [18]. While the number of 3D chromatin interaction experiments continue to increase, it is still not possible to perform chromatin interaction experiments for all types of cell and tissues. Therefore, computational approaches are urgently desired to complement experimental protocols.

Due to the limitations of experimental approaches, the number of available experimental data of EPIs is still limited. Several computational methods have been developed to predict EPIs. Depending on the type of the input data, computational methods can mainly be divided into two categories: DNA sequence-based methods and epigenomic data-based methods. For DNA sequence-based methods, PEP [19] and EP2vec [20] took advantage of natural language processing to learn the feature representation of DNA sequences, and SPEID [21] used convolutional neural network to learn the feature representation of DNA sequences. Recently, Zhuang et al [22] introduced a novel method to improve the prediction performance of EPIs by using the existing labeled data to pretrain a convolutional neural network (CNN), then adopting the training data from the cell line of interest to continue to train the CNN. Above these methods can accurately predict cell line-specific EPIs from genomic sequence, but they work well only when the training and testing data are from the same cell lines. For the epigenomic data-based methods, IM-PET [23], RIPPLE [24], TargetFinder [2], EpiTensor [25], and JEME [26] used many one-dimensional (1D) local chromatin states including but not limited to transcription factors (TFs) binding, histone modifications and chromatin accessibility signatures to predict EPIs. Though achieving acceptable performance, these models rely on labeled training data from the same cell line as the test data, which limits their usefulness for new cell lines. Generally speaking, high-resolution chromatin interactions experimental data are hard to get, and the cell-specific models have no acceptable generalization due to the specificity of EPIs. Therefore, how

to predict the EPIs of new cell line is an urgent problem. And the very intuitive idea is that to train a generic model which learns shared features between different cell lines, and then to predict the EPIs of the new cell line. However, this idea has its drawbacks. The feature distribution learned from the general model is different from that of a particular cell line that we care about. If the distribution of features learned from multiple cell lines is as similar as possible to the distribution of features in the cell line that we care about, then we can make more accurate predictions. Therefore, it is important to develop an effective method to make feature distribution as consistent as possible for transfer knowledge from other cell lines to a specific cell line that we care about.

It is well know that transfer learning (TL), an important branch of machine learning, is widely concerned in image recognition [27] and natural language processing [28], which focuses on the application of knowledge transfer onto new problems. Inspired by the work of [29], which suggested an adversarial neural network by gradient reversal layer (GRL) to fit the feature distribution of source domain and target domain for recognizing handwritten numbers, we proposed a novel method of SEPT to predict EPIs in the new cell line. There are no EPI labels information and only has the locations of enhancers and promoters in a particular cell line. The enhancer-promoter pairs of a particular cell line with no labels are defined as target domain, and labeled enhancer-promoter pairs from other cell lines are defined as source domain. Our goal is to learn the features from the source domain, and further reduce the data distribution differences between the target domain and the source domain through adversarial learning. First, we used convolution layers and long short-term memory (LSTM) layer to learn the features of EP pairs from the enhancer and promoter sequences. Then, adversarial neural network with GRL was used to reduce domain-specific features. GRL could reverse the direction of the gradient by multiplying the gradient with a negative constraints value. Finally, the trained model learns the EPI-related features from the source domain to predict the EPIs of the target domain.

SEPT is of great significance for three points. i) It could be used as an alternative

to the experimental methods, helping other tasks such as identifying mechanisms of SNPs from genome-wide association studies (GWAS) [30]. ii) It could reveal to what extent EPIs of one cell line could be recognized with the data from other cell lines. iii) It could improve our understanding of gene regulation and disease progression. To this end, we adopted two strategies: one is to combine different cell line data as the training data, and the other is to design the SEPT model with transfer learning [31] to transfer the informative features from the data of different cell lines to a new cell line. Experimental results show that SEPT has better performance than several other state-of-the-art methods. Model architecture analysis shows that LSTM layer and GRL layer are important for the EPIs prediction of across cell lines. Convolution kernels analysis shows that SEPT can effectively capture sequence features that determine EPIs.

2 Results

2.1 Comparison with other existing state-of-the-art methods

We first compared SEPT with other state-of-the-art methods of LS-SVM [32], SPEID [21] and RIPPLE [24]. SPEID and LS-SVM are the two sequence-based methods for predicting the DNA regulatory elements. LS-SVM [32] widely used the *k*-mer features, and did not take into account the interactions of high-order features. Because the LS-SVM [32] can only take one DNA sequence as an input sample, so we concatenated the sequences of the enhancer and promoter to train and test LS-SVM. SPEID [21] used deep learning to capture sequence features for predicting the cross-cell-line EPIs, while it lacks the ability to contain cell-specific information. RIPPLE [24] is a supervised model based on epigenomic data from ChIP-seq and DNase-seq experiments, and it only used five cell lines data to train the model, due to HMEC cell line lack of the epigenomic data. Our SEPT method simultaneously considers the source and target domain sequence information. For each test cell line, data from the other six cell lines were combined the training data to train SPEID [21], LS-SVM [32] and SEPT. The average results of SEPT, RIPPLE, SPEID and LS-SVM on seven test cell lines are shown in Figure 1, from which we can see that SEPT has the highest AUC values on

seven cell lines than RIPPLE, SPEID and LS-SVM. SEPT achieves 0.72, 0.76, 0.78, 0.77, 0.78, 0.73, and 0.76 for GM12878, HMEC, HUVEC, HeLa-S3, IMR90, K562, NHEK cell line, respectively. These results demonstrate that our SEPT can effectively predict the enhancer-promoter interactions.

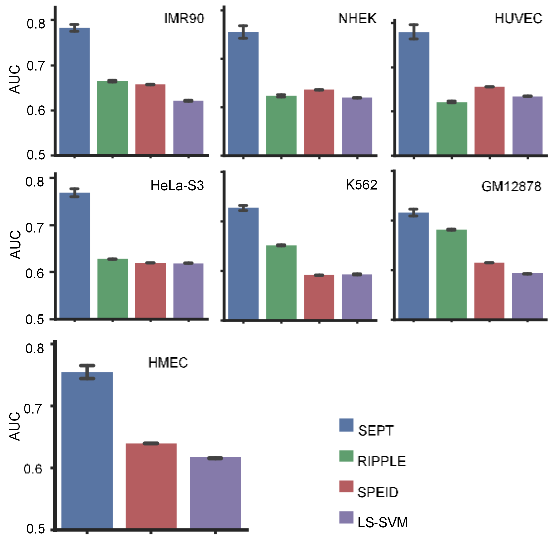

Fig. 1. Average AUC of SEPT, RIPPLE, SPEID and LS-SVM on seven test cell lines.

2.1 Influence of different neural network architectures in feature learning phase

We designed a series of computational network architectures in feature learning phase (Table S1) to investigate the impact of network structure. The first network architecture (namely BASE) includes only one convolutional layer in feature learning phase (Figure S1(a)). The second network architecture (namely BASE+LSTM) includes one convolutional layer and one LSTM layer in feature learning phase (Figure S1(b)). The third network architecture (namely BASE+FC) include one convolutional layer and one full connection layer in feature learning phase (Figure S1(c)). SEPT includes two convolutional layers and one LSTM layer in feature learning phase (Figure S1(d)). The grid search strategy was used to optimize the hyperparameters of four models in this work.

The average results of BASE, BASE+LSTM, BASE+FC and SEPT on seven test cell lines are shown in Figure 2, from which we can see that adding the full connection

layer (BASE+FC) and the LSTM layer (BASE+LSTM) in BASE model can improve the predictive performance of EPIs. Especially, the results of BASE+LSTM is better than that of BASE+FC, which indicates that the long-range dependency of DNA features such as motifs and other features can be captured by the LSTM layer. SEPT shows the best performance than BASE+LSTM, BASE+FC and BASE models, indicating that the first CNN layer learns the individual patterns in the sequences and the second CNN layer learns the high-order interactions between patterns. The high-order interactions may be commonalities among different cell lines, and deep neural network architectures can extract the high-order EPI features from DNA sequences.

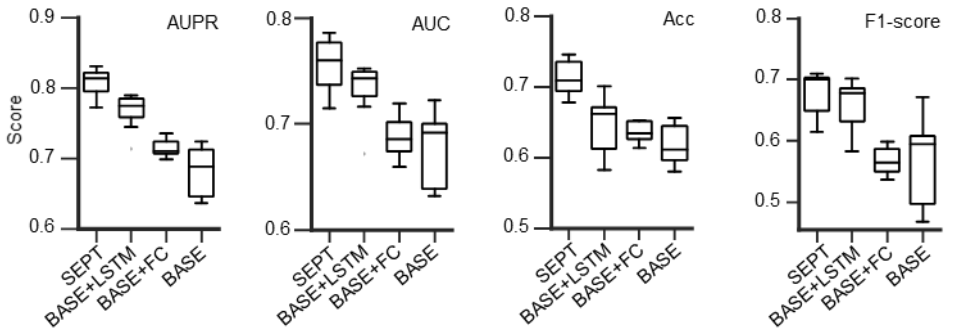

Fig. 2. Average results of the BASE, BASE+LSTM, BASE+FC and SEPT on seven test cell lines with the optimal hyperparameters.

2.3 Influence of domain adversarial operation

To validate the effectiveness of domain adversarial operation, we constructed another model (namely SEP), which has no the domain adversarial network architecture compared with SEPT. Since SEP has no the domain adversarial operation, the data of target domain cannot be used in SEP. That is, SEP just used the data of source domain in the training phase. Table 1 shows the average AUC results of SEPT and SEP by training model on one cell line data and test on another cell line data in running 10 times, from which we can see that AUC values of SEPT for each test cell line are higher than that of SEP, indicating that domain adversarial operation can effectively improve the predictive performance of EPIs in new cell lines. SEPT makes use of the EPIs information in other cell lines for recognizing the EPIs in new cell lines.

Table 1. Average AUC values of SEPT and SEP by training model on one cell line data and test on another cell line data in running 10 times. The row represents the labeled training cell lines and column represent the test cell lines.

Model	Test / Train	GM12878	HMEC	HUVEC	HeLa-S3	IMR90	K562	NHEK
SEPT	GM12878	*	0.63	0.68	0.66	0.67	0.62	0.62
	HMEC	0.59	*	0.57	0.61	0.67	0.56	0.59
	HUVEC	0.62	0.62	*	0.66	0.63	0.58	0.64
	HeLa-S3	0.61	0.62	0.69	*	0.66	0.63	0.66
	IMR90	0.58	0.65	0.62	0.59	*	0.56	0.62
	K562	0.64	0.61	0.66	0.66	0.62	*	0.64
	NHEK	0.62	0.63	0.66	0.67	0.66	0.59	*
SEP	GM12878	*	0.57	0.62	0.59	0.57	0.55	0.55
	HMEC	0.51	*	0.53	0.54	0.60	0.50	0.55
	HUVEC	0.59	0.55	*	0.58	0.57	0.56	0.55
	HeLa-S3	0.56	0.56	0.60	*	0.58	0.54	0.58
	IMR90	0.53	0.61	0.56	0.55	*	0.51	0.57
	K562	0.56	0.53	0.54	0.53	0.53	*	0.53
	NHEK	0.53	0.57	0.56	0.55	0.56	0.50	*

We also investigated the effectiveness of domain adversarial operation by training model on the six cell lines data and test on one other cell line data. The average AUC values of SEP and SEPT in running 10 times are shown in Table 2, from which we can see that average AUC of SEPT is higher 6% ~ 9% than that of SEP on seven test cell lines. The results in Table 2 show that the domain adversarial operation is also effective when using more cell lines data as the training data.

2.4 Results comparison of using different number of cell lines as the source domain data

To investigate the influence of different cell line number used in the source domain, we used different number of cell lines as the source domain data to predict the EPIs with SEPT. For selecting 1, 2, 3, 4, 5, 6 cell lines as the source domain data, each test cell line has the results of 6, 15, 20, 15, 6 and 1, respectively. The AUC results of using different number of cell lines as the source domain data to train SEPT are shown in

Figure 3. Each boxplot in Figure 3 represents the results of one test cell line across all combinations of source domain cell lines. To make the results more reliable, we repeated 10 times, that is, the AUC value of each test cell line is the mean of 10 running results. From figure 3, we can see that more cell lines as the source domain can achieves higher AUC than less cell lines as the source domain. Especially, when 6 cell lines are mixed as the source domain data to train the model, SEPT achieves the highest AUC values for every test cell line. With the increase of the cell lines number in source domain, AUC value is gradually larger. In addition, for different test cell line, the combination of appropriate cell lines as the source data can improve the performance of SEPT. These results show that using more existing labeled data of EPIs is helpful the prediction of EPIs in new cell line.

Table 2. Average AUC values of SEPT and SEP by training on the six cell lines data and test on one other cell line data in running 10 times.

Test cell line	Cell line(s) of source domain	Model	AUC
GM12878	HMEC, HUVEC, HeLa-S3, IMR90, K562, NHEK	SEP	0.65
		SEPT	0.72
HMEC	GM12878, HUVEC, HeLa-S3, IMR90, K562, NHEK	SEP	0.68
		SEPT	0.76
HUVEC	GM12878, HMEC, HeLa-S3, IMR90, K562, NHEK	SEP	0.69
		SEPT	0.78
HeLa-S3	GM12878, HMEC, HUVEC, IMR90, K562, NHEK	SEP	0.69
		SEPT	0.77
IMR90	GM12878, HMEC, HUVEC, HeLa-S3, K562, NHEK	SEP	0.72
		SEPT	0.78
K562	GM12878, HMEC, HUVEC, HeLa-S3, IMR90, NHEK	SEP	0.65
		SEPT	0.73
NHEK	GM12878, HMEC, HUVEC, HeLa-S3, IMR90, K562	SEP	0.69
		SEPT	0.76

2.5 Motifs identified by SEPT

As to investigate the motifs, we identified sequence features for each model by comparing patterns of the convolutional kernels in the first layer with sequence motifs from the database HOCOMOCO Human v11. We reconstructed the output of the first convolutional layer for each input sample sequence, then extracted the subsequence that

best match each kernel to compute the position frequency matrix (PFM) from the aligned sub sequences for each kernel. The motif comparison tool of Tomtom [33] was used to match these PFMs to known TF motifs. After obtaining the sequence motifs of the two models, we defined a metric of relative importance to measure the relative importance of motifs. The metric will consider the occurrence times of the same motif in both models and the ranks of the occurrence times of the motif in SEPT. If a motif appears more in SEPT and less in SEP, then the greater the effect of the motif in SEPT, the higher its relative importance score will be. The metric of relative importance is defined as: Relative importance $= (1 - N_{wo}/N_w)/Rank_w$. Here N_w and N_{wo} are the occurrence times of motif learned from SEPT and SEP, respectively; $Rank_w$ is the rank (in descending order) of the motif in the motif set of SEPT according to the occurrence times. The closer the value is to 1, the more important the motif is in SEPT. To avoid contingency, this process was repeated five times.

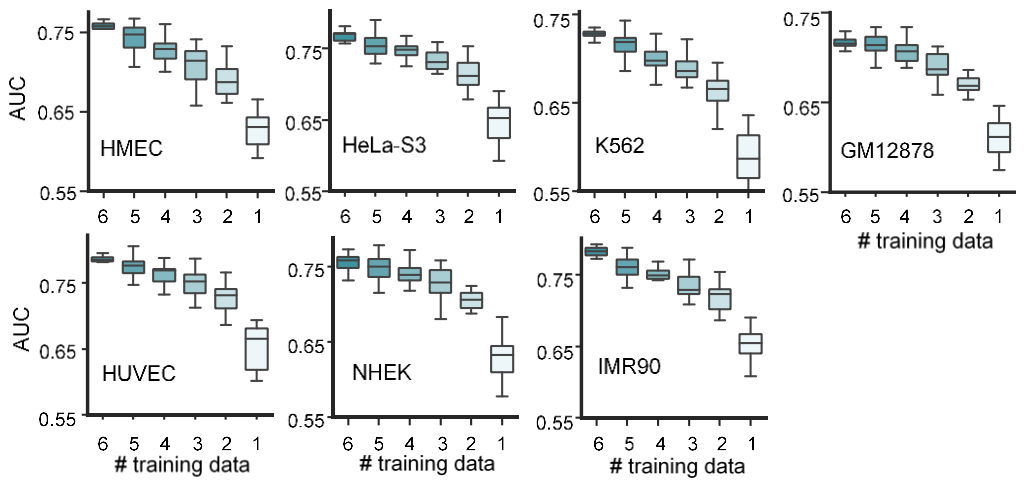

Fig. 3. AUC boxplots of different number of cell lines as the source domain in SEPT. The horizontal axis represents the number of cell lines in the source domain.

We found a set of potentially important transcription factor binding motifs by the transfer model. For each test cell line, top five import motifs involved in EPIs are show in Table S2. For different test cell lines, the models learned common important TF motifs such as ZNF563, THAP1, RXRA, and SP3, which are involved in many important processes, such as the transcriptional regulation and cell-cycle regulation. Interestingly, we found many of potentially important transcription factors that are associated with the corresponding cell lines (Table 3). For instance, MAFB motif learned by SEPT in the K562 cell line is associated with regulating lineage-specific hematopoiesis. This is consistent with the fact that K562 belong to a blood related cell line. NR4A1 motif learned by SEPT in GM12878 cell line was reported to play a role in the vascular response to injury, while ZNF341 motif learned by both SEPT and SEP in GM12878 cell line was reported to involve in the regulation of immune homeostasis. It is consistent with the fact that GM12878 is lymphoblastoid related cell lines. We also provided other motifs identified by both SEPT and SPEID (Table S3). These results show that our SEPT can learn important motifs, and these motifs are relevant to enhancer-promoter loops of a novel cell line.

Table 3. Examples of specific motifs identified by the first convolution layer of SEPT.

Cell line	Motif	Element	Brief description	Ref.
K562	MAFB, 	Promoter	It plays a pivotal role in regulating lineage-specific hematopoiesis.	[34]
K562	NR3C1, 	Enhancer	It affects inflammatory responses, cellular proliferation and differentiation in target tissues, and is involved in chromatin remodeling.	[35]
GM12878	NR4A1, 	Enhancer	It plays a role in the vascular response to injury.	[36]
GM12878	ZNF341, 	Promoter	It is involved in the regulation of immune homeostasis.	[37]
HeLa-S3	FOXK1, 	Promoter	It is involved in different processes such as glucose metabolism, aerobic glycolysis, muscle cell differentiation and autophagy.	[38]
HeLa-S3	BATF3, 	Promoter	It controls the differentiation of CD8+ thymic conventional dendritic cells in the immune system.	[39]
HUVEC	ZNF335, 	Enhancer	It controls the expression of genes involved in somatic development and regulates, for instance, lymphoblast proliferation.	[40]
IMR90	OLIG2, 	Enhancer	It is required for oligodendrocyte and motor neuron specification in the spinal cord.	[41]
NHEK	TFAP2C, 	Promoter	It is involved in important biological functions including proper eye, face, body wall, limb and neural tube development.	[42]
HMEC	ZNF317, 	Enhancer	It may play an important role in erythroid maturation and lymphoid proliferation.	[43]

3 Discussions

One important factor in SEPT is how to choose the labeled data of different cell lines as the source domain data. We chose different number of cell lines as the source domain data, and found that choosing all available labeled data of cell lines as the source domain data for training model can yield better performance than choosing partial cell lines data as the source domain. In addition, there is redundancy among different cell lines. Redundancy not only slows down model training, but also damages prediction performance. Thus, how to choose the different cell lines as the source domain data is important.

How to integrate other features such as histone modifications, chromatin accessibility and DNA shape into one model for predicting EPIs is also important factors. Although we only use sequence information in this work, the results of are still better than other methods which integrate the sequence and epigenomic information. Thus, if more information related to enhancers and promoters is integrated into our SEPT, it is hope that SEPT can significantly improve the performance of EPIs prediction.

Although SEPT can predict the potential EPIs in new cell line, it needs to provide the location of the enhancers and promoters in advance. Therefore, it is necessary to develop new methods for identifying the EPIs in specific cell line without enhancer and promoter location information.

4 Conclusions

Although some deep learning methods have been developed to predict EPIs within the same cells, they cannot achieve good performance for predicting EPIs in the unlabeled cell lines, due to lack of understanding of the interested cell lines. In this work, we proposed a transfer learning model to predict EPIs in new (or interested) cell lines. To better leverage the existing EPIs knowledge, we adopt the adversarial learning mechanism to learn useful features in the existing labeled cell lines and interesting unlabeled cell lines. Experiment results with the domain adversarial operation indicate

that it is helpful to predict EPIs in new cell lines. We expect that the model could learn informative features cross domains and reveal some commonalities (common TFs) between source and target domains. By learning commonalities between the source and target domains, SEPT outperforms other state-of-the-art methods for predicting EPIs in new cell lines.

Although SEPT can effectively predict EPIs in specific cell lines from enhancer and promoter sequences, it can be further improved by considering the following factors. i) SEPT just uses the sequence information of enhancers and promoters. Integrating other experimental data such as core histone modification ChIP-seq data or DNase-seq data can improve the performance of SEPT. ii) Each cell line is treated equally in the source domain, but the contribution of different cell lines should be different for the test cell line. Determining which cell lines should be used as training data is still needed to be explored, as more and more labeled data will become available. iii) Some EPIs maybe have the cell line specificity, while others are universal across many cell lines. Thus, different samples within the same cell line should have different contribution for cross-cell prediction. Assigning a proper weight to each sample can also improve the performance of SEPT.

5 Material and Methods

5.1 Data and preprocessing

We used the same Hi-C data as [2], and downloaded the Hi-C data of seven cell lines of K562 (mesoderm-lineage cells from a patient with leukemia), GM12878 (lymphoblastoid cells), HeLa-S3 (ectoderm-lineage cells from a patient with cervical cancer), HUVEC (umbilical vein endothelial cells), IMR90 (fetal lung fibroblasts), NHEK (epidermal keratinocytes) and HMEC (mammary epithelial cells) from Gene Expression Omnibus (GEO) GSE63525. The human reference genome hg19 was used to define the genomic locations. Promoters and activate enhancers in the first four cell lines were identified using segmentation-based annotations from both ENCODE Segway [44] and ChromHMM of Roadmap Epigenomics [45], only ChromHMM annotations were used in the other cell lines. Then, RNA-seq data from ENCODE were used to select activate promoters according to the rule of their mean FPKM >0.3 with

irreproducible discovery rate <0.1 for each cell line. The genome-wide Hi-C measurements were used to annotate all enhancer-promoter pairs as interacting or non-interacting in each cell type. For each enhancer-promoter pair, the distance between promoter and enhancer of the pair is more than 10kb and less than 2Mb [2]. To exclude the effect of distance on determining EPIs, interacting enhancer-promoter pairs were assigned to one bin (here, the total bin number sets to 5) based on quantile discretization of the distance between the enhancer and promoter. Random non-interacting pairs of active enhancers and promoters were assigned to their corresponding bin and then subsampled as the same number of positive samples within each bin. The subsampled non-interacting pairs were considered as negative samples. Table 4 gives the numbers of positive and negative pair in each cell line.

For each positive/negative sample, sequences of enhancers are extended or cut to 3kb flanking regions on location center of enhancers, and promoter are extended or cut to 2kb flanking regions on location center of promoters. One-hot coding format of enhancer and promoter sequence is used as input data of model.

We examined the overlapping number of positive EPIs between any two cell lines. For any two positive EPI pairs from any two cell lines, if the position of the two enhancers and the position of the two promoters both same, the two EPI pairs are considered to overlap. By removing redundancy, positive samples of different cell lines have very little overlap (Table S4).

Table 4. Number of enhancer-promoter interactions, enhancers and promoters in each cell line.

Cell line	#true EPIs	#all EPIs	#E	#P
IMR90	1254	2504	108996	5253
NHEK	1291	2571	144302	5254
HUVEC	1524	3044	65358	8180
HeLa-S3	1740	3480	102460	7794
K562	1977	3952	82806	8196
GM12878	2113	4226	100036	8453
HMEC	1342	2684	155328	5267

# denotes the number.

To compare with RIPPLE based on epigenetic data, we used the data from the Roadmap project for the six cell lines. Because we want to make EPI prediction across cell lines, we downloaded the peak files of 14 datasets that are measured in all six cell

lines. These datasets include CTCF, POLR2A, H2AZ, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1 and DNase-seq. An enhancer or promoter sample is represented as a binary vector in which each dimension corresponds to one of the epigenetic datasets. The feature vectors of enhancer and promoter are concatenated to represent each EPI pair. In addition, we also used other two features (i.e., the Pearson's correlation of the 14 signals associated with the enhancer-promoter pair, and the mRNA level of the gene associated with the promoter) to represent each EPI pair.

5.2 SEPT

Domain adaptation [28] is defined as that source domain and target domain share features and categories, but feature distribution is different. The source domain samples with rich information are used to improve the performance in target domain prediction. The source domain has abundant supervised labeling information, and the target domain has no or few labels. Because SEPT used the idea of domain adaptation, we therefore describe the input data in domain adaptation terms. As focusing on the task of EPIs prediction across cell lines, we assume that there is no labeled EPI information in test cell line, only the locations of enhancers and promoters are provided. So, we can utilize the abundant supervised labeling information of other cell lines (called source domain) to improve the performance of EPIs prediction in new cell lines without labeled EPI information (called target domain).

An overview of SEPT is shown in Figure 4(a). For predicting the EPIs in cell line #A, unlike the existing two methods which extract the specific features (in Figure 4(b)) or shared features (in Figure 4(c)) of cell lines, SEPT uses the rich information of cell lines #B and #C to extract the features that are relevant to the EPIs in cell line #A by using the transfer learning (TL). As shown in Figure 4(a), we used GRL to design the domain discriminator. GRL reverses the gradient direction during the back propagation, but it does nothing when forward propagation. Obviously, without GRL, SEPT will gradually learn the features related to domain in the training process. With the GRL, the model learns the features that cannot tell whether the data is from the source domain

(labeled data of mixed cell lines #B, #C) or the target domain (unlabeled data of cell lines #A) in training process. SEPT simultaneously trains two classifiers of the main label classifier and the domain discriminator. These two classifiers share feature extractor layers. SEPT mainly includes feature extractor, domain discriminator and EPI predictor.

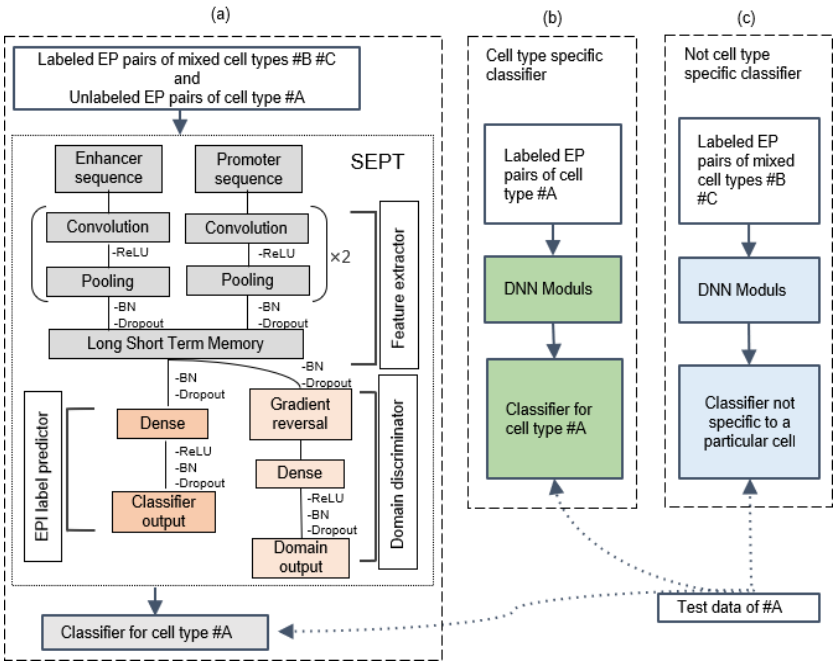

Fig. 4. Deep neural network-based methods for predicting EPIs. (a) SEPT architecture. SEPT uses the feature extractor to learn EPIs-related features from the mixed cell line data, meanwhile the domain discriminator with the transfer learning mechanism is adopted to remove the cell-line-specific features and retain the features independent of cell line. EPI label predictor identifies the EPIs in new cell line based on the learned features. (b) Previous model trains a model on specific cell line data, in which the training and test data are both from same cell line. (c) Mixed cell line data are used to train a general model for predicting EPIs in a new cell line.

Feature extractor consists of two convolution layers, two max-pooling layers, two dropout layers, and one recurrent long short-term memory (LSTM) layer. Domain discriminator consists of the GRL, one dense layer, one dropout layer, and the output. EPIs predictor consists of one dense layer, one dropout layer, and the output. Since informative features may differ between enhancers and promoters, we use two convolution layers, rectified linear unit (ReLU) and max-pooling layers for enhancers

and promoters, respectively. Thus, the inputs are two one-hot matrixes to represent enhancer and promoter sequence, respectively. Because large number of kernels can sufficiently extract the features, and motif features of DNA sequences are short than 40bp, so each convolution layer consists of 300 kernels with length 40. Max-pooling layer reduces the output dimension with pool length 20, stride 20. The outputs of the two branches are concatenated into one tensor, which is input to the dropout layer with dropout rates of 0.25. The dropout layer randomly selects partial input data to next layer to avoid overfitting. The recurrent LSTM layer is used to extract feature combinations of the two branches, and the output dimension of LSTM is 100. For domain discriminator, the output of LSTM layer feeds into GRL, and a dense layer maps the learned distributed features to the domain label space. It contains 50 units with ReLU activations. The output feeds into a sigmoid unit to predict the domain probability after dropout layer with dropout rates of 0.5. For EPI predictor, the output of LSTM layer feeds into dense layer, which further maps the learned distributed features to the sample label space. It contains 100 units with ReLU activations. The output feeds into a sigmoid unit to predict the probability after dropout layer with dropout rates of 0.5.

SEPT Model training

We trained SEPT for 80 epochs with mini-batches of 64 samples by back-propagation. In the training phase, source domain data were used to train the feature extractor and the EPI predictor, and both source and target domain data were used to train the feature extractor and the domain discriminator. SEPT seeks to minimize the loss of EPIs label and domain discriminator. Binary cross-entropy loss function for both EPIs label predictor and domain discriminator is used, which is minimized by stochastic gradient descent (SGD) with initialized learning rate 0.001. In view of the two optimization objectives, SEPT learns a discriminative representation for EPI prediction and indistinguishable representation for domain prediction. The objective function of the SEPT is defined as follows:

$$E(\theta_f, \theta_y, \theta_d) = \sum_{i=1}^{N} L_y\left(G_y(G_f(x_i;\theta_f);\theta_y), y_i\right) - \lambda \sum_{i=1}^{M} L_d\left(G_d(G_f(x_i;\theta_f);\theta_d), d_i\right) \quad (1)$$
$$= \sum_{i=1}^{N} L_y^i\left(\theta_f, \theta_y\right) - \lambda \sum_{i=1}^{M} L_d^i\left(\theta_f, \theta_d\right)$$

Here, L_y is the loss of EPIs label predictor, L_d is the loss of domain discriminator, G_f is a mapping that maps the input x to a feature vector, G_y is a mapping that maps the feature vector to the EPIs label, G_d is a mapping that maps the feature vector to the domain label, x_i is the i-th sample, θ_f is the parameters of mapping G_f, θ_y is the parameters of mapping G_y, θ_d is the parameters of mapping G_d, y_i is the EPI label of the i-th sample, d_i is the domain label of the i-th sample, N is the number of labeled EPI training samples, M is the number of unlabeled EPI training samples but they have the domain labels, and λ is a constant that controls the tradeoff between two objectives.

It is a problem of minimax optimization, that is, we attempts to seek a saddle point of the functional $E(\theta_f, \theta_y, \theta_d)$ that is delivered by parameters $\widehat{\theta}_f$, $\widehat{\theta}_y$, and $\widehat{\theta}_d$. At the saddle point, the loss of EPI label predictor and domain discriminator is minimized. The loss of EPI label predictor is minimized by the feature mapping parameter θ_f, while the loss of domain discriminator is maximized by θ_f on account of the GRL that changes the sign of the gradient during the back-propagation. In the end, SEPT learns the features that are discriminative and domain-invariant. The features learned from the cell lines with label information (source domain) are effective for the new cell lines (target domain).

The training procedure of SEPT can be described as follows: i) Randomly separating the dataset of target domain into approximate equal two parts, one as the training data in which each sample has domain label but no EPI label, and the other as the testing data. ii) Mixing the data from other six cell lines and randomly shuffling the data. The mixed data are used as source domain dataset in which each sample has both domain and EPIs labels. iii) Training SEPT with the source domain data and target

domain data. iv) Evaluating the performance of SEPT with the test data in target domain.

All the experiments are based on Python by using the scikit-learn machine learning library [46] and Keras framework (https://keras.io/) with Tensorflow as back-end [47].

5.3 Evaluation metrics

We adopted the metrics of Accuracy, Precision, Recall, F1 score, AUC and AUPR to assess the performance of SEPT. These metrics are defined respectively as the following [48-51].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (5)$$

where *TP* is the number of correctly predicted EPIs, *TN* is the number of correctly predicted non-EPIs, *FP* is the number of incorrectly predicted EPIs and *FN* is the number of incorrectly predicted non-EPIs. AUC is the area under the receiver operating characteristic (ROC) curve which is the plot of the true-positive rate (i.e., sensitivity) as a function of false-positive rate (i.e., 1-specificity) based on various thresholds. AUPR is the area under the precision-recall curve which is the plot of the precision as a function of recall based on various thresholds.

Abbreviations

EPIs: Enhancer-promoter interactions; AUC: Area under the receiver operating curves; 3D: Three-dimensional; 3C: Chromosome conformation capture-based approach; CNN: Convolutional neural network; GWAS: Genome-wide association studies; TL: Transfer learning; GRL: Gradient reversal layer; LSTM: Long short-term memory; ReLU:

Rectified linear unit; SGD: Stochastic gradient descent; AUPR: Area under the precision-recall curves.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The related source codes and datasets are available at https://github.com/NWPU-903PR/SEPT.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the National Natural Science Foundation of China (No. 61873202 to SWZ; No. 11661141019 and 61621003 to SZ]; the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [No. XDB13040600], the National Ten Thousand Talent Program for Young Top-notch Talents, the Key Research Program of the Chinese Academy of Sciences, [No. KFZD-SW-219] and CAS Frontier Science Research Key Project for Top Young Scientist [No. QYZDB-SSW-SYS008]. The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

FJ designed and performed the experiments, and wrote the initial manuscript. Both FJ and SZ designed the methods. SWZ and SZ revised the manuscript. All authors participated in the definition of the process and approved the final manuscript.

Acknowledgments

The authors would like to thank our group members for their valuable suggestions.

Authors' information

Fang Jing: Email: jingnpu@foxmail.com; Address: 127 West Youyi Road, Xi'an, Shaanxi, 710072, China

Shao-Wu Zhang: Email: zhangsw@nwpu.edu.cn; Address: 127 West Youyi Road, Xi'an, Shaanxi, 710072, China

Shihua Zhang: Email: zsh@amss.ac.cn; Address: 55 Zhongguancun East Road, Beijing, 10090, China.

References

1. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem. 2009;107(1):30-9.
2. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016; 48(5):488-96.
3. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014; 15(4):272-86.
4. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461(7261):199-205.
5. Van Steensel B, Dekker J. Genomics tools for unraveling chromosome architecture. Nat Biotechnol. 2010; 28(10):1089-95.
6. Bickmore WA, van Steensel B. Genome architecture: domain organization of interphase chromosomes. Cell. 2013; 152(6):1270-84.
7. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. Cell. 2016; 164(6):1110-21.
8. Rowley MJ, Corces VG. The three-dimensional genome: principles and roles of long-distance interactions. Curr Opin Cell Biol. 2016; 40:8-14.
9. Achinger-Kawecka J, Clark SJ. Disruption of the 3D cancer genome blueprint. Epigenomics. 2017; 9(1):47-55.
10. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016; 538(7624):265-69.
11. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa

IS, Beaudry JL, Puviindran V. FTO obesity variant circuitry and adipocyte browning in humans. New Engl J Med. 2015; 373(10):895-907.

12. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015; 161(5):1012-25.

13. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. science. 2002; 295(5558):1306-11.

14. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science. 2009; 326(5950):289-93.

15. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376-80.

16. De Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013; 502(7472):499-506.

17. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7):1665-80.

18. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T. An atlas of active enhancers across human cell lines and tissues. Nature. 2014; 507(7493):455-61.

19. Yang Y, Zhang R, Singh S, Ma J. Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics. 2017; 33(14):i252-i260.

20. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics. 2018; 19(2):84. doi:10.1186/s12864-018-4459-6

21. Singh S, Yang Y, Póczos B, Ma J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol. 2019; 7(2):122-37.

22. Zhuang Z, Shen X, Pan W. A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data. Bioinformatics. 2019; 17(35):2899-906.

525 23. He B, Chen C, Teng L, Tan K. Global view of enhancer–promoter interactome in human
526 cells. Proceedings of the National Academy of Sciences. 2014; 111(21):E2191-E2199.
527 24. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R. A
528 predictive modeling approach for cell line-specific long-range regulatory interactions.
529 Nucleic Acids Res. 2015; 43(18):8694-712.
530 25. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L,
531 Wang W. Constructing 3D interaction maps from 1D epigenomes. Nat Commun. 2016;
532 7:10812. doi:10.1038/ncomms10812
533 26. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MT, Cheng C, Fan X, Gerstein M.
534 Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues
535 and cell lines. Nat Genet. 2017; 49(10):1428-36.
536 27. Sun B, Saenko K: Deep coral. Correlation alignment for deep domain adaptation. In:
537 European Conference on Computer Vision. 2016; Springer: 443-50.
538 28. Abdelwahab M, Busso C. Domain adversarial for acoustic emotion recognition.
539 IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018; 26(12):2423-
540 35.
541 29. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M,
542 Lempitsky V. Domain-adversarial training of neural networks. The Journal of Machine
543 Learning Research. 2016; 17(1):2096-30.
544 30. Guo H, Ahmed M, Zhang F, Yao CQ, Li S, Liang Y, Hua J, Soares F, Sun Y, Langstein J.
545 Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to
546 prostate cancer. Nat Genet. 2016; 48(10):1142-50.
547 31. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data
548 engineering. 2009; 22(10):1345-59.
549 32. Dongwon L. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics. 2016;
550 32(14): 2196-98.
551 33. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. et al. Quantifying similarity between
552 motifs. Genome Biol. 2007; 8, R24. doi: https://doi.org/10.1186/gb-2007-8-2-r24
553 34. Eychène, Alain, Rocques N , Pouponnot C. A new MAFia in cancer. Nature Reviews Cancer.
554 2008; 8(9):683-93.

555 35. Fryer C, Archer T. Chromatin remodelling by the glucocorticoid receptor requires the
556 BRG1 complex. Nature. 1998; 393(6680):88-91.
557 36. Papac-Milicevic N, Breuss JM, Zaujec J, et al. The interferon stimulated gene 12 inactivates
558 vasculoprotective functions of NR4A nuclear receptors. Circulation Research. 2012;
559 110(8):e50-e63.
560 37. Stefanie FJ, Hartberger JM, Manfred F, et al. ZNF341 controls STAT3 expression and
561 thereby immunocompetence. ence Immunology. 2018; 3(24):eaat4941.
562 doi:10.1126/sciimmunol.aat4941
563 38. Bowman CJ, Ayer DE, Dynlacht BD. Foxk proteins repress the initiation of starvation-
564 induced atrophy and autophagy programs. Nature Cell Biology. 2014; 16(12):1202-14.
565 39. Bower KE, Fritz JM, Mcguire KL. Transcriptional repression of MMP-1 by p21SNFT and
566 reduced in vitro invasiveness of hepatocarcinoma cells. Oncogene. 2004; 23(54):8805-14.
567 40. Yang YJ, Baltus AE, Mathew RS, et al. Microcephaly gene links trithorax and REST/NRSF
568 to control neural stem cell proliferation and differentiation. Cell. 2012; 151(5): 1097-112.
569 41. Beyer CA, Cabanela ME, Berquist TH. Unilateral facet dislocations and fracture-
570 dislocations of the cervical spine. The Bone & Joint Journal. 1992; 73(6):977-81.
571 42. Bamforth SD, Bragança J, Eloranta JJ, et al. Cardiac malformations, adrenal agenesis,
572 neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nat
573 Genet. 2001; 29(4):469-74.
574 43. Takashima H, Nishio H, Wakao H, et al. Molecular cloning and characterization of a
575 KRAB-containing zinc finger protein, ZNF317, and its isoforms. Biochemical and
576 Biophysical Research Communications. 2001; 288(4): 771-9.
577 44. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL,
578 Barrell D, Zadissa A, Searle S. GENCODE: the reference human genome annotation for
579 The ENCODE Project. Genome Res. 2012; 22(9):1760-74.
580 45. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P,
581 Zhang Z, Wang J, Ziller MJ. Integrative analysis of 111 reference human epigenomes.
582 Nature. 2015; 518(7539):317-30.
583 46. Swami A , Jain R. Scikit-learn: Machine learning in Python. Journal of machine learning
584 research. 2011; 12(10):2825-30.

47. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. Tensorflow: A system for large-scale machine learning. arXiv preprint arXiv:1603.04467.

48. Jing F, Zhang S, Cao Z, Zhang S. An integrative framework for combining sequence and epigenomic data to predict transcription factor binding sites using deep learning. IEEE/ACM transactions on computational biology and bioinformatics. 2019; 1-1. doi: 10.1109/TCBB.2019.2901789

49. Fan XN, Zhang SW. LPI-BLS: Predicting lncRNA-protein interactions with a broad learning system-based stacked ensemble classifier. Neurocomputing. 2019; 370:88-93.

50. Zhang SW，Zhang XX, Fan XN, Li WN. LPI-CNNCP: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. Analytical Biochemistry. 2020; 601: 113767.

51. Fan XX, Zhang SW, Zhang SY and Ni JJ. lncRNA_Mdeep: An Alignment-Free Predictor for Distinguishing Long Non-Coding RNAs from Protein-Coding Transcripts by Multimodal Deep Learning. Int. J. Mol. Sci. 2020; 21: 5222.