Subject Section

Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network

Wanwen Zeng[1], Yong Wang[2, 3*] and Rui Jiang[1, *]

[1]MOE Key Laboratory of Bioinformatics; Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China

[2]CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100080, China

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

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Interactions among cis-regulatory elements such as enhancers and promoters are main driving forces shaping context-specific chromatin structure and gene expression. Although there have been computational methods for predicting gene expression from genomic and epigenomic information, most of them neglect long-range enhancer-promoter interactions, due to the difficulty in precisely linking regulatory enhancers to target genes. Recently, HiChIP, a novel high-throughput experimental approach, has generated comprehensive data on high-resolution interactions between promoters and distal enhancers. Moreover, plenty of studies suggest that deep learning achieves state-of-the-art performance in epigenomic signal prediction, and thus promoting the understanding of regulatory elements. In consideration of these two factors, we integrate proximal promoter sequences and HiChIP distal enhancer-promoter interactions to accurately predict gene expression.

Results: We propose DeepExpression, a densely connected convolutional neural network, to predict gene expression using both promoter sequences and enhancer-promoter interactions. We demonstrate that our model consistently outperforms baseline methods, not only in the classification of binary gene expression status but also in regression of continuous gene expression levels, in both cross-validation experiments and cross-cell line predictions. We show that the sequential promoter information is more informative than the experimental enhancer information; meanwhile, the enhancer-promoter interactions within ±100 kbp around the TSS of a gene are most beneficial. We finally visualize motifs in both promoter and enhancer regions and show the match of identified sequence signatures with known motifs. We expect to see a wide spectrum of applications using HiChIP data in deciphering the mechanism of gene regulation.

Availability: DeepExpression is freely available at https://github.com/wanwenzeng/DeepExpression.

Contact: ruijiang@tsinghua.edu.cn, ywang@amss.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene regulation, as one of the fundamental problems in biology, explains how different types of cells in the human body emerge from the identical

information encoded by the genome (Ozbudak, *et al*., 2002). The transcription of a gene is an extremely intricate process that requires a complex set of interactions among *trans* proteins and *cis* DNA sequences (Maston, *et al.,* 2006). The regulation of this process is accomplished in large part by promoters and enhancers, which are DNA sequences containing multiple binding sites for a variety of transcription factors (TFs) (Yao, *et al*., 2015). Enhancers can activate transcription independent of their location, distance, or orientation with respect to the promoters of genes (Heinz, *et al*., 2013). Therefore, ever since the emergence of high-throughput experiments for quantifying gene expression, computational biologists have long been interested in how well gene expression can be inferred by TFs and regulatory elements (Rockman and Kruglyak, 2006), for deciphering the mechanism of gene regulation.

In computational studies of gene regulation (Lee and Young, 2013), various experimental data related to one-dimensional (1D) epigenomic signals, including TFs binding (Li, *et al.*, 2014), histones modification (Karlic, *et al.*, 2010) and chromatin accessibility (Duren, *et al.,* 2017), are taken as features to predict gene expression. These methods mainly fall into two categories. The first class of methods predict whether gene expression level is high or low under a binary classification formulation. For example, DeepChrome (Singh, *et al.*, 2016) used five histone markers in promoter regions with a convolutional neural network (CNN) to predict gene expression. The second class of methods infer continuous gene expression levels under a regression framework and thus can provide quantitative predictions. For example, Ouyang *et al.* used ChIP-Seq data of 12 TFs in mouse embryonic stem cells (mESC) with a linear regression model to predict gene expression (Ouyang, *et al*., 2009). Karlić *et al.* collected nineteen histones modification in promoter regions in mESC to regress gene expression (Karlic, *et al.*, 2010). Dong *et al.* used twelve histone modification markers and chromatin accessibility in promoter regions with a two-step model for gene expression prediction (Dong, *et al.*, 2012). However, these methods have some limitations. First, they never explicitly take enhancers and three-dimensional (3D) enhancer-promoter interactions into consideration thus far, probably due to both the difficulty in accurately linking enhancers to their target genes and the uncertainty of how strong these interactions will affect the gene expression (Mora, *et al.*, 2015). Second, most of these methods collect several specific histone markers' and TFs' ChIP-seq data. Therefore, they are limited by the heavy data demand and hard to be generalized across different cell lines. Third, these methods either predict gene expression as binary classification problem or infer gene expression level as the regression problem. Few of them could handle both situations.

The recent development of HiChIP (Mumbach, *et al*., 2016), a high-throughput experimental technique for sensitive and efficient analysis of protein-centric chromosome conformation, holds the promise to capture chromatin loops with high sensitivity and specificity. Compared with Hi-C (Belton, *et al.,* 2012) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Li, *et al.,* 2014), HiChIP is able to measure protein-centric chromatin conformation in a rapid, efficient, technically simplified and high-resolution way. Among existing HiChIP studies, two of them stand out to show the ability of HiChIP in identifying enhancer-promoter interactions. First, Mumbach *et al.* evaluated H3K27ac, an enhancer- and promoter-associated mark, as a candidate factor to selectively interrogate enhancer-promoter interactions across the genome (Mumbach, *et al.*, 2017). Second, Weintraub *et al.* found the binding of YY1 activated enhancers and promoter-proximal elements and formed dimers to facilitate the interaction of these DNA elements (Weintraub, *et al.,* 2017). Therefore, HiChIP experiments of H3K27ac and YY1 have been developed to identify high-confidence 3D chromatin loops located around enhancer-promoter interactions. These datasets provide valuable raw and general materials for us to study the regulatory functions of enhancer-promoter interactions on gene expression.

Besides the rapid progress in biological experiments, recently, deep learning techniques have achieved the state-of-the-art performance on many bioinformatics applications such as regulatory site identification (Alipanahi, *et al.*, 2015) and biomedical image classification (Krizhevsky, *et al.,* 2017). A deep learning model automatically learns a complex nonlinear function that maps inputs onto outputs, eliminating the need to use hand-crafted features or rules. In bioinformatics, CNNs have been used to predict regulatory elements (Min, *et al.*, 2017), chromatin accessibility (Liu, *et al.*, 2018) and epigenetic states of a DNA fragment (Min, *et al.*, 2017; Zhou and Troyanskaya, 2015), as well as to explain functional implications of genetic variants (Zhou and Troyanskaya, 2015).

Inspired by the promising HiChIP experiments and the advanced deep learning techniques, we introduce DeepExpression, a deep learning framework to model gene expression, with consideration of enhancers, promoters, and their interactions. For distal enhancers, we adopt the state-of-the-art high-resolution 3D HiChIP experiments as features. For proximal promoters, we apply a recently developed deep learning model, called densely connected convolution neural networks, to extract epigenomic features in promoter regions. Cross-validation and cross-cell line prediction experiments show that DeepExpression consistently outperforms several baseline methods not only in the classification of binary gene expression status but also in the regression of continuous gene expression levels. Model ablation analysis indicates that both the promoter information and enhancer information are informative for gene expression prediction. Furthermore, through a visualization strategy, we show that DeepExpression successfully captures sequence motifs in both promoter and enhancer regions, which are matched in the JASPAR database (Khan, *et al.*, 2018).

2 Methods

2.1 Data collection and preprocessing

We collected HiChIP data of H3K27ac for mESC and identified corresponding RNA-seq data (Weintraub, *et al.,* 2017). We collected HiChIP data of YY1 for the human HCT116, Jurkat, and K562 cell lines (Weintraub*, et al.,* 2017) and identified corresponding RNA-seq data from the ENCODE project (Consortium, 2012). We extracted DNA fragments of 2,000 base pairs (bp) around Transcription Start Site (TSS) of a gene as its promoter region. The summary of the data is shown in Supplementary Table 1.

We followed the preprocessing pipeline described in (Weintraub, *et al*., 2017) to deal with RNA-seq data and HiChIP data (Supplementary Methods). Gene expression levels were calculated by applying a logarithmic transformation of base 10 to gene-level counts, after adding a pseudocount of α ($\alpha = 1$), and then quantile normalized across samples. For HiChIP data, we followed Weintraub *et al*. to divide whole genome DNA sequences into bins of length 5 kbp. To adjust for different sequencing depths, we divide interaction counts n_{ijk} for interaction between bin i and bin j in sample k by the total read count of the sample N_k and then scale the result by multiplying the minimum read count of all samples N. After this procedure, the raw count n_{ijk} for inteaction between bin i and bin j of sample k was converted into a normalized read count $\tilde{n}_{ijk} = {}^{n_{ijk}}/_{N_k} \times N$.

Normalized interaction counts $\tilde{n}_{ijk}$ from a total of n replicate samples were averaged and then logarithmic transformed with base 2 after adding

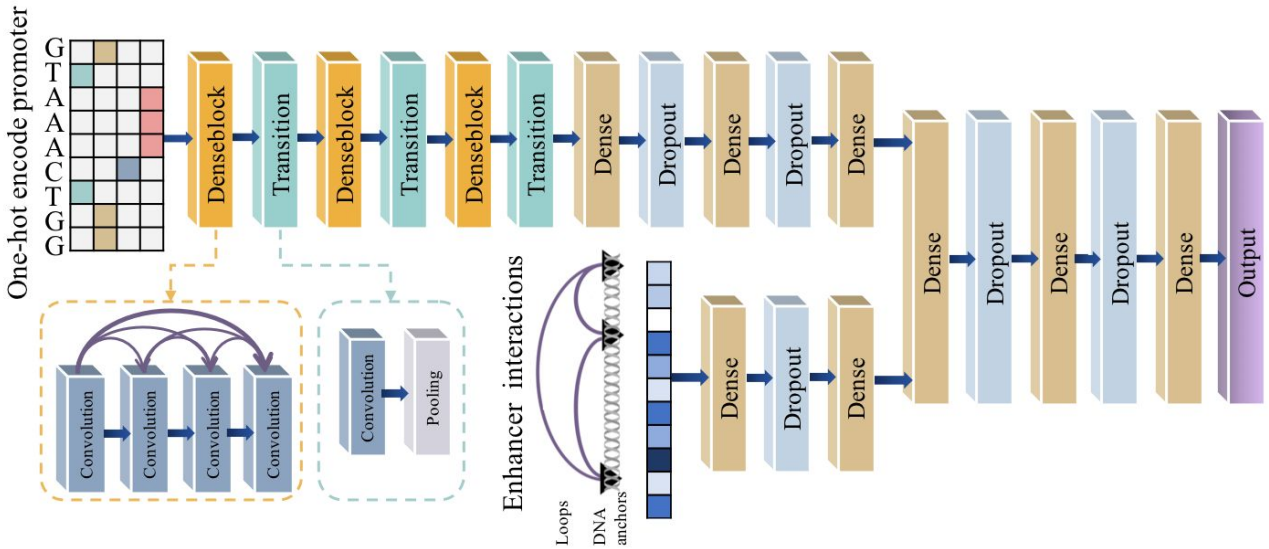

Figure 1. The graphical illustration of DeepExpression. First, a sequential promoter module is pre-trained to extract features from the input promoter regions. Second, an experimental HiChIP enhancer-promoter interactions module is adopted to fine-tune DeepExpression. Finally, a joint module integrates the outputs of the promoter and enhancer modules to predict the gene expression.

a pseudocount of β ($\beta = 1$) to characterize the interaction affinity of an interaction in a cell line. The value of HiChIP interaction signal between bin *i* and bin *j* is therefore

$$h_{ij} = \log_2\left(1 + \frac{1}{n}\sum_{k=1}^{n}\tilde{n}_{ijk}\right).$$

For each TSS in a specific bin_p, we extracted HiChIP interactions signals ±1,000 kbp (bin_{p-200}, bin_{p-199},…, bin_{p+200}) around bin_p. Each bin includes adjacent positions of 5 kbp flanking bin *p*. Then for each gene, the HiChIP interaction feature is a 400-dimensional real value vector and each dimension represents HiChIP long-range enhancer-promoter interactions signal h_{pq}between the specific bin_q, $q \in (p-200, p-199, \ldots, p+200)$ and TSS-located bin_p.

2.2 Design of DeepExpression

As illustrated in Figure 1, DeepExpression consists of three modules. First, a proximal promoter module is used to extract features from DNA sequences in promoter regions. Second, a distal enhancer-promoter interaction module is used to extract features of HiChiP enhancer-promoter interactions signals. Finally, a joint module integrates the outputs of the above two modules to produce a predicted gene expression signal.

2.2.1 Proximal promoter module as a densely connected convolutional neural network

The proximal promoter module consists of three main components: a one-hot encoding input layer, three densely connected convolution blocks and three fully connected layers.

The one-hot encoding layer converts a DNA fragment into a numerical representation for downstream processing. It encodes the nucleotide in each position as a four-dimensional one-hot binary vector, in which each element represents one type of nucleotide: A, C, G, and T. The encoding layer then concatenates the binary vectors into a 4-by-2,000 binary matrix, to represent the whole 2,000-bp target sequence.

The densely connected convolution blocks automatically extract features for an encoded DNA fragment. Recent advances in deep learning have shown that a classical convolutional neural network usually has hundreds of thousands of parameters involved, and thus often results in severe overfitting problem on tasks with small datasets (Srivastava, *et al.*, 2014). Hence, a densely connected convolution network (Huang, *et al.,* 2017) was proposed to utilize parameters more efficiently and avoid the overfitting problem, which connects all layers directly with each other. As schematically illustrated in Figure 1, in a block consists of L (L = 4) convolution layers, the input of the l-th layer is the concatenation of the feature-maps produced by all the preceding layers $0,\ldots, l-1$, as

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$$

where H_l denotes the concatenation operation. Meanwhile, the feature-maps of the l-th layer are passed on to all $L - l$ subsequent layers. This introduces $L(L+1)/2$ connections in an L-layer network, instead of just L, as in a traditional architecture of convolutional neural networks.

The convolution operation could be formulated as

$$Conv(X)_{ik} = Relu\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{mn}^{k} x_{i+m,n}\right)$$

where X is the input matrix, M is the size of the sliding window, N is the number of input channels, and $W^k = (w_{mn}^k)_{M \times N}$ is the weight matrix of the k-th convolution kernel with size $M \times N$. In the first convolution layer, N is equal to 4. This first convolution process is equivalent to scanning a position weight matrix (PWM) across the target sequence. In the other convolution layers, N is equal to the total number of convolutional kernels of all the preceding layers. The convolution layer then applies the rectified linear unit (ReLU) nonlinear function as

$$Relu(x) = max(0, x)$$

The pooling layer computes the maximum in each of the non-overlapping windows of size M, providing invariance to local shifts and reducing the number of parameters.

$$Pool(X)_{ik} = max(x_{iM,k}, x_{iM+1,k}, \ldots, x_{iM+M,k})$$

Three fully connected layers with 80, 40, and 40 units, respectively, performs linear transformations of the outputs of the previous layer, and applies the rectified linear unit nonlinear function. Finally, the proximal promoter module transforms each sequential input to a 40-dimensional real vector.

2.2.2 Distal enhancer-promoter interaction module as a feedforward neural network

The distal enhancer-promoter interaction module receives 400-dimensional HiChIP enhancer-promoter interactions signals as input. It uses two fully connected layers with 80 and 40 units to transform the 400-dimensional numeric enhancer-promoter interaction strength input to a 40-dimensional real-valued feature vector.

2.2.3 Joint expression prediction module as a feedforward neural network

The joint module integrates different features from both the proximal promoter and distal enhancer modules to predict gene expression. We merge outputs of these two modules to form a feedforward neural network. For binary classification model, we use a softmax function to produce a probability output as $f_i(z) = e^{z_i}/\sum_j e^{z_j}$, where $f_i(z)$ is the predicted probability that the input gene belongs to class i. For continuous regression model, we modify the output layer by replacing the softmax layer with a linear transformation layer.

2.3 Training of DeepExpression

The detailed selection of different network structure and hyper-parameters combinations in enhancer-promoter interactions module are provided in Supplementary Methods. After choosing the network structure, we trained the proposed model in a two-step approach. First, in a pre-training step, we optimized the cross-entropy loss in the classification model or the mean squared error loss in the regression model, without the distal enhancer-promoter interaction module. We used the RMSprop (Hinton, *et al*.) optimizer with a batch size of 4 and used dropout with a 0.5 dropout rate for model regularization. We also applied the early stopping strategy (Erhan, *et al*., 2010) with the maximum number of iterations set as 60, and it would stop training after 5 epochs of unimproved loss on the validation set. We denote the model trained in this step as DeepExpression-seq. Second, in a fine-tuning step (Liu, *et al*., 2015), we incorporated the enhancer-promoter interaction module before the joint output module. For fine-tuning, we initialized the promoter module using the DeepExpression-seq model and then optimized loss function on the whole network. We implemented our method by Keras, a deep learning library for Theano and Tensorflow. We used Theano as the backend, while the Tensorflow backend also generated very close results through our testing. We used the high-performance NVIDIA Tesla K80 GPU for model training.

We also tried to train both modules simultaneously, instead of using pre-training and fine-tuning steps separately. We compare the performance of these two strategies in Supplementary Table 2. We could observe that, by fine-tuning, DeepExpression could be trained more effectively and needed fewer iterations to converge. Besides, using a two-step training strategy, for cell lines without HiChIP experiments, we could still derive a DeepExpreesion-seq model only using sequential information.

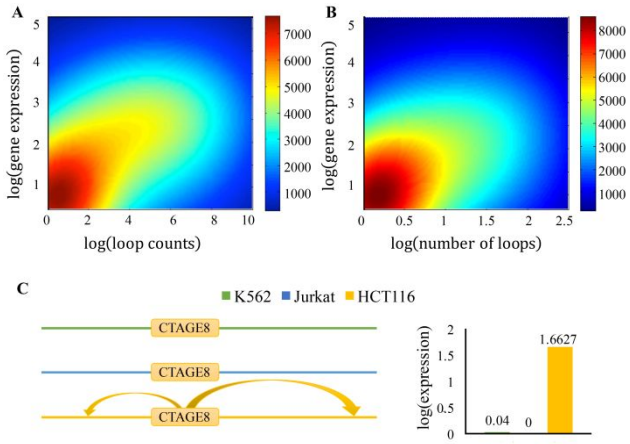

Figure 2. A) Scatter plot of HiChIP loop counts and gene expression in mESC (PCC: 0.623). The color bar on the right indicates the density of the scatter plot. B) Scatter plot of the number of HiChIP loops and gene expression in mESC (PCC: 0.583). The color bar on the right indicates the density of the scatter plot. C) Visualization of the HiChIP loops of CTAG8, the expression of CTAG8 in K562, Jurkat, and HCT116 cell lines respectively.

2.4 Evaluation of DeepExpression

We adopted multi-fold cross-validation experiments to evaluate our model. Taking 10-fold as an example, we randomly split each dataset into ten strictly non-overlapping groups. For each run, we used nine groups to train our model and the rest one as for testing. Data of the nine groups was further split as a training set and a validation set with ratio 0.8:0.1. The training set was used to adjust weights in the network, and the validation set was used for early stopping to avoid overfitting.

To validate the generalization ability of DeepExpression, we further adopt cross-cell line validation and independent validation. For cross-cell line validation, we trained our model in one single cell line and predicted in another cell line. For independent validation, for each cell line, we used the independently trained regression models of the other cell lines to make predictions, and then averaged over the resulting regression values to obtain a final regression value for a test gene.

2.5 Comparison with baseline machine learning models

To evaluate the performance of DeepExpression, we implemented three baseline methods for classification (logistic regression, SVM, and random forest) and three methods for regression (linear regression, Lasso regression, and random forest regression). All the methods took both sequential and experimental data as input in accord with DeepExpression.

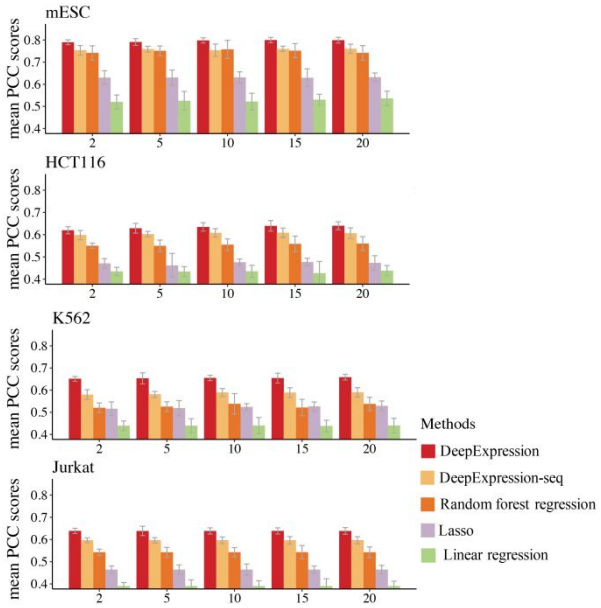

Figure 3. The regression performance measured in PCC at different *K*-folds cross-validation experiments (*K*=5, 10, 15, 20).

For sequence data, we split the sequence of a DNA fragment into k-mers in a sliding window fashion with stride of 1 bp ($k = 6$). For example, for a promoter with 2,000bp length, we will have 1,995 6-mers. Then we represent each enhancer sequence with corresponding counts for each 6-mer using a vector with dimension equal to 4,096 (4^6). For experimental data, we take signals (loop counts in HiChIP experiments) for bins to form an input vector with dimension equal to 400. Then we combine these two vectors together and run the baseline methods. We performed an internal 10-fold cross-validation experiment for model selection among regularization parameter and hyper-parameter configurations. The detail of model selection is provided in Supplementary Methods. For those gene expression prediction methods that we mentioned in the introduction section, we could not compare our method with them in all our cell lines, since some of their features are not available in some cell lines. Specifically, DeepChrome used 5 histone markers and Dong *et al*. used 12 histone modification markers and chromatin accessibility. For these two methods, K562 was the only overlapped cell line with our data. Ouyang et al. used ChIP-Seq data of 12 TFs and only trained their model in mESC cells. Karlić *et al*. collected 19 histones modification in promoter regions and also only trained model in mESC. For these two methods, mESC was the only overlapped cell line with our data. Therefore, we re-implemented these methods using the same cross-validation strategy as DeepExpression in the corresponding matched cell line.

3 Results

3.1 HiChIP enhancer-promoter interactions are discriminative features for predicting gene expression

Since no previous studies have shown contributions of chromatin interactions to gene expression in a quantitative way, we first devised and tested the ability of HiChIP enhancer-promoter interactions to discriminate gene expression levels from the mESC cell line. For each gene, we first simply extracted all the loops interacting with it, then summed all the loop counts, which measure the strength of an interaction, and drew the scatter plot of loop counts and gene expression. Figure 2A shows that the Pearson Correlation Coefficient (PCC) between gene expression levels and total HiChIP loop counts is up to 0.623. We also counted the number of loops of each gene and drew the scatter plot of the number of loops and gene expression. Figure 2B shows that correlation between gene expression and the number of HiChIP loops is also high with PCC around 0.583. We also tested the ability of HiChIP in human K562 cell line (Supplementary Figure 1) and got similar results. We could conclude that these HiChIP signals are positively correlated with gene expressions and provide informative features to predict gene expressions. We further show a simple example of the relationship between gene expression and HiChIP signal in three human cell lines. CTAGE8 is an important paralog of CTAGE4, a gene associated with Cutaneous T Cell Lymphoma (Usener, *et al.*, 2003). We find that CTAGE8 only expresses in the HCT116 cell line and almost has no expression in K562 and Jurkat. From HiChIP data, no enhancer-CTAGE8 interactions are detected in K562 and Jurkat while two active enhancer-CTAGE8 interactions are found in HCT116.

From the above analysis, we could draw the conclusion that cell-type specific enhancer-promoter interactions obtained from HiChIP indeed provide useful regulatory information on gene expression, and thus are discriminative features for predicting gene expression.

3.2 DeepExpression accurately models gene expression

To recover the level of gene expression, we modeled gene expression as the response variable and built DeepExpression regression model. We compared the performance of DeepExpression regression with three baseline methods, including linear regression, Lasso, and random forest regression. We also compared with our alternative model, DeepExpression-seq, which discarded the HiChIP experimental data integration module and regressed gene expression using only DNA sequence information. We systematically evaluated the performance of DeepExpression in capturing gene expression codes via a series of carefully designed multi-fold cross-validation experiments. We repeated the cross-validation experiments for the different number of folds, evaluated the performance of each method using PCC scores, and reported the regression performance in Figure 3.

As shown in Figure 3, our method consistently outperforms all the baseline methods. For example, in the 10-fold cross-validation experiment for mESC, the PCCs of our method are on average 0.0404, 0.1669, and 0.2763 higher than random forest regression, Lasso, and linear regression, respectively. It is also worth noting that the DeepExpression-seq model is also superior to the three baseline methods and performs more stably. For example, in the 10-fold cross-validation experiment for mESC, the PCCs of DeepExpression-seq are on average 0.0020, 0.1227, and 0.2321 higher than random forest, Lasso, and linear regression, respectively. Our method also demonstrates much stronger robustness than the baseline methods in the regression task. With variances of PCCs calculated for cross-validation experiments of different folds for each cell line, one-sided paired-sample Wilcoxon rank sum tests as described in the previous section consistently suggest that our method achieves significantly smaller variance than a baseline method (p-value < 3.6e-8 for all the three methods).

Besides, we also modeled gene expression as a binary classification problem (Supplementary Methods) and reported the performance of DeepExpression classification model in Supplementary Figure 2. Furthermore, we compared the performance of DeepExpression with

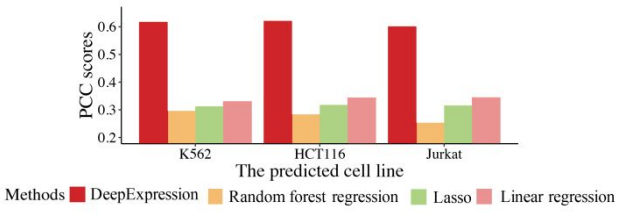

Figure 4. The independent prediction performance measured in PCC at different testing cell lines. Given a cell line, we used DeepExpression models trained on all the other available cell lines to predict gene expression, and then averaged over the predictions to obtain the final prediction result of the gene.

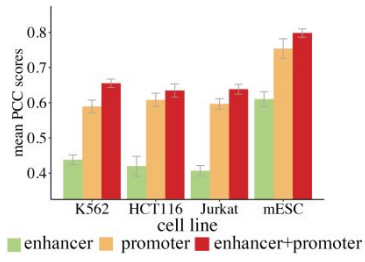

Figure 5. Contributions of sequential promoter and experimental enhancer features. We performed a model ablation analysis by repeating the cross-validation experiments with the enhancer-promoter interaction modules excluded to evaluate the contribution.

other gene expression prediction methods in specific cell lines and found that our method also achieved higher performance than other methods (Supplementary Table 3). In summary, the superior performance of our method in all cell lines, and in both regression and classification tasks, indicates the powerful prediction ability of DeepExpression.

3.3 Cross-cell line and independent prediction

A HiChIP experiment provides a means of measuring how strong an enhancer regulates a target gene in a cell line. We wonder whether it could be possible to impute the expression of a gene in a cell line with the incorporation of HiChIP experimental data of other cell lines. To simulate this scenario, we performed a series of experiments for cross-cell line prediction and independent prediction. For cross-cell line prediction, we trained DeepExpression in a specific cell line and predicted in another cell line. For independent prediction, given a cell line, we used DeepExpression models trained on all the other available cell lines to predict gene expression, and then averaged over the predictions to obtain the final prediction result of the gene.

We first used datasets of three human cell lines to demonstrate the ability of our method to regress gene expression in a cross-cell line manner. As shown in Figure 4, this independent predicting strategy is consistently superior to the three baseline methods. In detail, the PCCs of our method are on average 0.33, 0.30, and 0.27 higher than random forest, Lasso, and linear regression, respectively. We also showed the cross-cell line prediction result in Supplementary Table 4, where DeepExpression achieved higher performance than baseline models. It is worth noting that DeepExpression-seq performs well in within-cell line prediction tasks, its performance is much lower in the cross-cell line predictions tasks (Supplementary Table 5), because the expression values are different in different cell lines but the promoter sequence inputs remain the same. Adding the HiChIP information, DeepExpression well performs in both independent and cross-cell line prediction tasks.

These results suggest that by combining information of both promoters and enhancer-promoter interactions, we might be capable of predicting gene expression across different cell types. Notably, taking housekeeping genes (Eisenberg and Levanon, 2013) to evaluate cross-cell line performance, we find that DeepExpression achieved higher performance for predicting housekeeping genes than all the genes (Supplementary Table 6). We expect to train DeepExpression in more and more cell lines incorporating HiChIP enhancer-promoter interactions data, and consequently we could predict the gene expression for a new cell line that has not been studied yet. More importantly, we expect to learn the comprehensive and general gene regulation mechanisms with enhancer regulation across different cell lines.

3.4 Contributions of sequential and experimental features

The distal enhancer-promoter interaction module incorporates experimental HiChIP long-range enhancer-promoter interaction information into our methods. To prove that the experimental data is informative, we performed a model ablation analysis by repeating the cross-validation experiments with the enhancer-promoter interaction modules excluded. In a similar way, we excluded the promoter module to evaluate its contribution.

As shown in Figure 5, there are evident differences in the contributions of the proximal promoter and distal enhancer-promoter interaction module. Taking mESC as an example, after removing the promoter module, the mean PCC decreases by 23.64%. When removing the enhancer-promoter module, however, the mean PCC drops by 8.39%. Obviously, promoter sequences provide more information than enhancer-promoter interaction data to predict gene expression. We speculate that there are two reasons accounting for this phenomenon. First, we incorporate HiChIP enhancer-promoter interactions using a fine-tuning way, while those primitive feedforward networks might not capture all the information in HiChIP data. Second, since HiChIP is a newly developed experimental technique, there is no formal pipeline to process HiChIP data, and thus we might lose some information during the processing procedure.

We also computed the contribution of enhancer-promoter interactions and promoters at different distances. To evaluate the contribution, we carried out the sensitivity analysis for different lengths of the HiChIP experimental input region and different length of the promoter sequential input region. The detailed results are shown in Supplementary Table 7 and 8. We could conclude that the information of HiChIP enhancer-promoter interactions is most beneficial from those within ±100 kbp around the TSS of a gene while the information of promoter sequential information is most beneficial from TSS ±1000 bp. However, we could still conclude that using sequential promoter data and experimental enhancer jointly effectively improves the performance and play important roles in predicting gene expression.

3.5 DeepExpression recovers TF binding motifs in promoter and enhancer regions

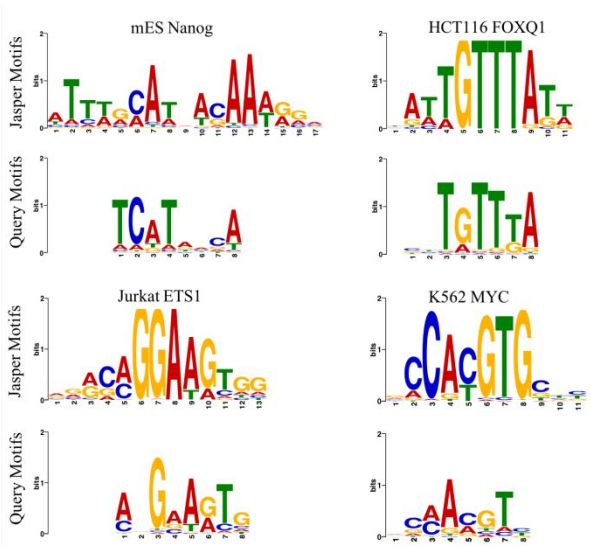

Figure 7. Motif visualization by enhancer regions in each cell line.

The formation of enhancer-promoter interactions needs some structural proteins binding such as cohesin and other TFs binding (Weintraub, *et al*., 2017). One assumption is that some important TFs/structure proteins will bind to both enhancer regions and promoter regions (Supplementary Figure 3) to help to form the three-dimensional enhancer-promoter interactions (Kim and Shiekhattar, 2015). To evaluate the assumption and demonstrate the interpretability of our model, we identified motifs learned from promoter and enhancer regions separately and we found some motifs could be identified in both promoter and enhancer regions. For promoter regions, we identified motifs in the first convolution layer of DeepExpression using the strategy described in Supplementary Methods. For enhancer regions, we applied CisModule (Zhou and Wong, 2004) to visualize motifs learned from enhancer sequences in HiChIP data (Supplementary Methods). We then compared these motifs with known Vertebrates motifs. Using motif comparison tool TomTom with significant *E*-value threshold 0.05, we matched about 65% (83/128) of motifs learned in promoters to known motifs in different cell lines, as shown in Figure 6, while we matched about 92% (22/24) of motifs from HiChiP interactions in Figure 7. Moreover, the four distinguished motifs learned from promoter regions are also learned by the CisModule.

To name a few, we showed some motifs learned both in promoter and enhancer regions: in mESC, DeepExpression recovers Nanog, a transcription factor involved in embryonic stem cell proliferation, renewal, and pluripotency (Han, *et al.,* 2008). In HCT116, DeepExpression recovers FOXQ1, a member of the FOX gene, which is involved in embryonic development, cell cycle regulation, tissue-specific gene expression, cell signaling, and tumorigenesis (Qiao, *et al.,* 2011). In Jurkat, DeepExpression discovers ETS1, which functions either as transcriptional activators or repressors and are involved in stem cell development, cell senescence and death, and tumorigenesis (Thomas, *et al.,* 1995). In K562, DeepExpression discovers MYC that plays a role in cell cycle progression, apoptosis and cellular transformation (Gomez-Casares, *et al.,* 2013). Amplification of this gene is frequently observed in numerous human cancers. Translocations involving this gene are associated with Burkitt lymphoma and multiple myeloma in human patients (Ceballos, *et al.,* 2000).

The consistency of TFs discovered in promoter regions and enhancer regions explains why using both promoters and enhancers features jointly could improve DeepExpression performance. The powerful learning ability of DeepExpression could not only help us find potential TFs binding in a specific cell line, but also guide us to find novel motifs which have not been discovered by experiments yet. Furthermore, the motif relevance in promoter regions and enhancer regions will be modeled explicitly in future version of DeepExpression.

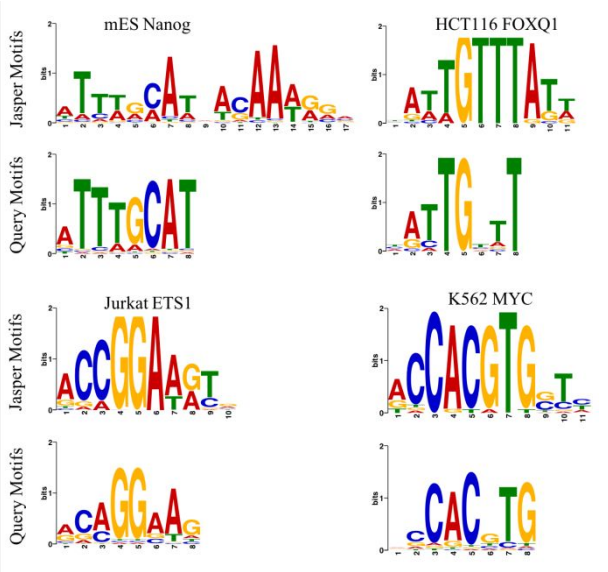

Figure 6. Motif visualization by promoter regions in each cell line.

4 Discussion

We have introduced a deep learning framework named DeepExpression to integrate DNA sequence information and enhancer-promoter interaction data for modeling gene expression. Through comprehensive validation experiments, we have shown that DeepExpression is superior to baseline methods in different cell lines and different species, is capable of making cross-cell line predictions, and is interpretable in extracted features.

DeepExpression is distinct from other methods for predicting gene expression in the following aspects. First, we adopt novel state-of-the-art 3D HiChIP experimental features while existing methods only use 1D features such as histone modification and chromatin accessibility (Shu, *et al*., 2011). HiChIP defines the high-resolution landscape of enhancer-promoter regulation. Many complex features of the 3D enhancer connectome cannot simply be predicted from 1D data, demonstrating that it is necessary to employ these features. Second, we combine promoters and enhancer features together to model gene expression. Enhancers and promoters are the most important cis-regulatory elements and have a huge impact on gene expression. Taking these two types of features into account, we could better model gene expression.

It is worth noting that DeepExpression performs better in stem cells than other differentiated cell lines. One reason for this phenomenon may be that the gene expression pattern in stem cell is much simpler than other differentiated cell lines, especially K562, Jurkat, and HCT116 are all cancer cell lines. The expression patterns might be dysregulated in these cell lines so it is much more difficult to predict the gene expressions.

Nevertheless, our work can be further improved in several aspects. First, the adaptation of an embedding representation of DNA sequences instead of using the one-hot encoding may also benefit the prediction accuracy (Min, *et al.*, 2017). Second, since we have shown that the first convolutional layer could capture motif information, researchers may use our model to learn the complex grammar of TF binding in specific cell lines. In addition, one can also explore interactions of motifs in higher convolutional layers. Third, our deep learning framework can possibly be adapted for the integration of other 3D functional elements interactions in the genome, including but not limited to silencers, repressors and insulators (Raab and Kamakaka, 2010). Fourth, we could better model the motif information located in promoter and enhancer regions. Through section 3.5, we could *de novo* discover important motifs in promoters and enhancers respectively. We could combine these motifs information in a unified framework to model gene expression. Fifth, we can try to perform cross-species prediction to measure the 'cross-species gap' (Normand, *et al.*, 2018). Sixth, we could incorporate image-like HiChIP data using the densely connected CNN to better extract the experimental features. Seventh, since 1D features have shown to effective to predict gene expression, we can further integrate 1D features such as chromatin accessibility from ATAC-seq data. Last but not least, we look forward to deciphering the enhancer-promoter interactions regulatory mechanism across species.

Acknowledgements

Rui Jiang is a RONG professor at the Institute for Data Science, Tsinghua University.

Funding

This work was partially supported by the National Key Research and Development Program of China (No. 2018YFC0910404), the National Natural Science Foundation of China (Nos. 61873141, 61721003, 61573207), and the Tsinghua-Fuzhou Institute for Data Technology.

Conflict of Interest: none declared.

References

Alipanahi, B., *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 2015;33(8):831-+.

Belton, J.M., *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 2012;58(3):268-276.

Ceballos, E., *et al.* c-Myc antagonizes the effect of p53 on apoptosis and p21(WAF1) transactivation in K562 leukemia cells. *Oncogene* 2000;19(18):2194-2204.

Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.

Dong, X., *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome biology* 2012;13(9):R53.

Duren, Z., *et al.* Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci U S A* 2017;114(25):E4914-E4923.

Eisenberg, E. and Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet* 2013;29(10):569-574.

Erhan, D., *et al.* Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 2010;11(Feb):625-660.

Gomez-Casares, M.T., *et al.* MYC antagonizes the differentiation induced by imatinib in chronic myeloid leukemia cells through downregulation of p27(KIP1). *Oncogene* 2013;32(17):2239-2246.

Han, M.K., *et al.* SIRT1 regulates apoptosis and Nanog expression in mouse embryonic stem cells by controlling p53 subcellular localization. *Cell Stem Cell* 2008;2(3):241-251.

Heinz, S., *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* 2013;503(7477):487-+.

Hinton, G., Srivastava, N. and Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Huang, G., *et al.* Densely connected convolutional networks. In, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 3.

Karlic, R., *et al.* Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* 2010;107(7):2926-2931.

Khan, A., *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* 2018;46(D1):D1284.

Kim, T.-K. and Shiekhattar, R.J.C. Architectural and functional commonalities between enhancers and promoters. 2015;162(5):948-959.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun Acm* 2017;60(6):84-90.

Lee, T.I. and Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* 2013;152(6):1237-1251.

Li, G., *et al.* Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* 2014;15 Suppl 12:S11.

Li, Y., Liang, M. and Zhang, Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol* 2014;10(10):e1003908.

Liu, Q., *et al.* Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics* 2018;34(5):732-738.

Liu, Z., *et al.* Deep learning face attributes in the wild. In, *Proceedings of the IEEE international conference on computer vision*. 2015. p. 3730-3738.

Maston, G.A., Evans, S.K. and Green, M.R. Transcriptional regulatory elements in the human genome. *Annu Rev Genom Hum G* 2006;7:29-59.

Min, X., *et al.* Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics* 2017;18(Suppl 13):478.

Min, X., *et al.* Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics* 2017;33(14):I92-I101.

Mora, A., *et al.* In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in bioinformatics* 2015;17(6):980-995.

Mumbach, M.R., *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016;13(11):919-922.

Mumbach, M.R., *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* 2017;49(11):1602-1612.

Normand, R., *et al.* Found In Translation: a machine learning model for mouse-to-human inference. *Nat Methods* 2018;15(12):1067-1073.

Ouyang, Z., Zhou, Q. and Wong, W.H. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A* 2009;106(51):21521-21526.

Ozbudak, E.M., *et al.* Regulation of noise in the expression of a single gene. *Nat Genet* 2002;31(1):69-73.

Qiao, Y., *et al.* FOXQ1 regulates epithelial-mesenchymal transition in human cancers. *Cancer Res* 2011;71(8):3076-3086.

Raab, J.R. and Kamakaka, R.T. OPINION Insulators and promoters: closer than we think. *Nat Rev Genet* 2010;11(6):439-446.

Rockman, M.V. and Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* 2006;7(11):862-872.

Shu, W.J., *et al.* Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res* 2011;39(17):7428-7443.

Singh, R., *et al.* DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 2016;32(17):i639-i648.

Srivastava, N., *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 2014;15:1929-1958.

Thomas, R.S., *et al.* Ets1 Transactivates the Human Gm-Csf Promoter in Jurkat T-Cells Stimulated with Pma and Ionomycin. *Oncogene* 1995;11(10):2135-2143.

Usener, D., *et al.* cTAGE: A cutaneous T cell lymphoma associated antigen family with tumor-specific splicing. *Journal of Investigative Dermatology* 2003;121(1):198-206.

Weintraub, A.S., *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 2017;171(7):1573-1588. e1528.

Weintraub, A.S., *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 2017;171(7):1573-1588 e1528.

Yao, L.J., Be rman, B.P. and Farnham, P.J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit Rev Biochem Mol* 2015;50(6):550-573.

Zhou, J. and Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12(10):931-934.

Zhou, Q. and Wong, W.H. CisModule: De novo discovery of' cis-regulatory modules by hierarchical mixture modeling. *P Natl Acad Sci USA* 2004;101(33):12114-12119.