Bioinformatics, 33(13), 2017, 1930–1936
doi: 10.1093/bioinformatics/btx105
Advance Access Publication Date: 17 February 2017
Original Paper

Sequence analysis

BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone

Bite Yang[1,†], Feng Liu[1,2,†], Chao Ren[1], Zhangyi Ouyang[1], Ziwei Xie[3], Xiaochen Bo[1,*] and Wenjie Shu[1,*]

[1]Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing 100850, [2]Department of Information, The 188th Hospital of Chaozhou, Chaozhou 521000 and [3]Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on November 8, 2016; revised on February 14, 2017; editorial decision on February 15, 2017; accepted on February 16, 2017

Abstract

Motivation: Enhancer elements are noncoding stretches of DNA that play key roles in controlling gene expression programmes. Despite major efforts to develop accurate enhancer prediction methods, identifying enhancer sequences continues to be a challenge in the annotation of mammalian genomes. One of the major issues is the lack of large, sufficiently comprehensive and experimentally validated enhancers for humans or other species. Thus, the development of computational methods based on limited experimentally validated enhancers and deciphering the transcriptional regulatory code encoded in the enhancer sequences is urgent.
Results: We present a deep-learning-based hybrid architecture, BiRen, which predicts enhancers using the DNA sequence alone. Our results demonstrate that BiRen can learn common enhancer patterns directly from the DNA sequence and exhibits superior accuracy, robustness and generalizability in enhancer prediction relative to other state-of-the-art enhancer predictors based on sequence characteristics. Our BiRen will enable researchers to acquire a deeper understanding of the regulatory code of enhancer sequences.
Availability and Implementation: Our BiRen method can be freely accessed at https://github.com/wenjiegroup/BiRen.
Contact: shuwj@bmi.ac.cn or boxc@bmi.ac.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Enhancers are distal *cis*-acting DNA regulatory elements that play key roles in gene expression in a time- or cell-line-specific manner (Bulger and Groudine, 2010, 2011; Calo and Wysocka, 2013; Ong and Corces, 2011). Understanding the properties, genomic targets and regulatory activities of enhancers is currently an area of great interest, given the increasing appreciation of their importance in development (Stathopoulos and Levine, 2005), cell identity (Hnisz *et al.*, 2013; Loven *et al.*, 2013; Whyte *et al.*, 2013), phenotypic diversity (Levine and Tjian, 2003), evolution and human disease (Visel *et al.*, 2009). Given the absence of common sequence features, the distal location from their regulated targets and their high cell type/tissue specificity, the accurate identification of enhancers remains a significant challenge in the annotation of mammalian genomes.

In recent years, the advent of deep sequencing has enabled the development of a large variety of computational methods for enhancer identification that complement experimental techniques by

integrating different data types derived from different data sources. Based on the available data sources, enhancer identification methods can be grouped into three categories in a conceptually simple manner (Kleftogiannis *et al.*, 2015a,b); however, different computational methods rely on an integration of different datasets/features and/or a combination of supervised and unsupervised components. The first category includes bioinformatics approaches that identify enhancers using epigenetic profiles, such as histone markers derived from ChIP-seq, DNase I hypersensitivity sites (DHSs) and/or transcription factor-binding sites (TFBSs), mainly through clustering and unsupervised learning techniques. The second category of methods reformulates the enhancer identification problem as a binary classification task by discriminating enhancer regions from non-enhancer (negative set) regions using supervised machine learning techniques, such as support vector machines (SVMs) (Fernandez and Miranda-Saavedra, 2012; Kleftogiannis *et al.*, 2015a,b), artificial neural networks (ANNs) (Firpi *et al.*, 2010), decision trees (DTs) (Lu *et al.*, 2015), random forests (RFs) (Rajagopal *et al.*, 2013), probabilistic graphical models (PGMs) (Ernst and Kellis, 2012; Hoffman *et al.*, 2012) and, more recently, deep learning (Liu *et al.*, 2016). The third category represents a variety of bioinformatics methods based on high-resolution data derived from enhancer testing and screening methods to detect and test enhancers in human, mouse, flies and yeast (Shlyueva *et al.*, 2014). However, despite major efforts to develop accurate enhancer prediction methods, these bioinformatics methods still encounter numerous issues in addition to technical problems, such as the class-imbalance problem, over-fitting issues, tuning of model parameters and poor generalization ability. One major obstacle is the lack of a large, sufficiently comprehensive and experimentally validated enhancer set for humans or other species. Thus, the development of computational methods based on limited experimentally validated enhancers and deciphering the transcriptional regulatory code encoded in enhancer sequences is urgent.

In this study, we developed a deep-learning-based hybrid architecture, named BiRen, that integrates the sequence encoding and representation power of a convolutional neural network (CNN) and the superior capacity for handling the long-term dependency of long DNA sequences of a gated recurrent unit (GRU)-based bidirectional recurrent neural network (BRNN) to accurately identify enhancers using the DNA sequence alone. BiRen was trained with limited experimentally validated enhancer elements derived from the VISTA Enhancer Browser (Visel *et al.*, 2007) that exhibit gene enhancer activity, as assessed in transgenic mice. We demonstrate that BiRen directly learns regulatory code from genomic sequences and illustrates superior identification accuracy, robustness of overcoming noise data, and generalization to other species for enhancer predictions relative to two state-of-the-art methods based on sequence characteristics such as motifs or *k*-mers. Our BiRen will provide researchers with a deeper understanding of the regulatory code of enhancer sequences.

2 Materials and methods

2.1 Datasets

In total, 1747 and 567 experimentally validated human and mouse noncoding elements with gene enhancer activity, as assessed in transgenic mice, were collected from the VISTA Enhancer Browser (Visel *et al.*, 2007), respectively. For the human noncoding fragments, 900 elements were defined as POSITIVE enhancers that exhibited reproducible expression in the same structure in at least three independent transgenic embryos, whereas 847 elements were defined as NEGATIVE enhancers that exhibited no reproducible expression in any structure in at least three different embryos. For the mouse noncoding fragments, 322 and 245 elements were defined as POSITIVE and NEGATIVE enhancers, respectively. Evolutionarily conserved features were taken from the vertebrate phastCons44way track (Siepel *et al.*, 2005) in the UCSC Genome Browser (Goldman *et al.*, 2015). Additionally, DHSs of 125 human cells and H3K27 ac-binding sites of 86 human cells/tissues were derived from the ENCODE project (ENCODE Project Consortium, 2012) and the Roadmap Epigenomics Project (Kundaje *et al.*, 2015), respectively.

2.2 Construction of the enhancer and non-enhancer sets

We constructed the enhancer set using the experimentally validated human and mouse noncoding fragments in the VISTA Enhancer Browser (Visel *et al.*, 2007). In total, 900 POSITIVE human enhancers and 322 POSITIVE mouse enhancers were collected as positive sets in the human and mouse genomes, respectively. The length of the validated human enhancers ranged from 428 to 8061 bp with a median of 1334 bp, and the length of validated mouse enhancers ranged from 330 to 5099 bp with a median of 1573 bp. The non-enhancer set contained random genomic loci not annotated as promoters or enhancers (10 times the size of the enhancer set). Promoters were defined as 2-kb regions centred on transcriptional start sites (TSSs) of protein-coding genes, and random genomic loci were generated with an equivalent length distribution as enhancer regions.

To assess the robustness of the ability of BiRen to overcome the noise associated with a false-positive enhancer training set, we combined the POSITIVE and NEGATIVE enhancers in the VISTA Enhancer Browser to construct new enhancer sets in the human and mouse genomes that contained 9900 and 3542 elements, respectively. Correspondingly, we generated new non-enhancer sets that contained random genomic loci not annotated as promoters or enhancers (10 times the size of the new enhancer set).

2.3 Encoding the DNA sequence by a CNN

A CNN is a well-known deep learning architecture that has been extensively applied to computer vision, natural language processing, speech recognition and other artificial intelligence research fields (Cornu and Milner, 2015; Lawrence *et al.*, 1997; Meng *et al.*, 2014). The basic components of a CNN include convolutional, pooling and fully connected layers. The convolutional layer aims to extract and represent the local information of raw features by several feature maps and kernels (weight matrices). The pooling layer aims to compress the resolution of the feature maps to achieve spatial invariance. After several convolution and pooling operations, there may be one or more fully connected layers to perform high-level reasoning. The output of the last fully connected layer is fed to an output layer. For a classifier or regression task, softmax regression is commonly used, as it generates a well-formed probability distribution of the outputs (Krizhevsky *et al.*, 2012).

The recent method DeepSEA (Zhou and Troyanskaya, 2015) successfully applies CNNs to sequence-based problems in genomics. The entire deep convolutional network uses three convolution layers with 320, 480 and 960 kernels and corresponding three max pooling layers. A fully connected layer of 925 neurons and 919 outputs are connected after the convolution and pooling layers. After training with large-scale chromatin-profiling data from the ENCODE project (ENCODE Project Consortium, 2012), DeepSEA can accurately predict chromatin features for the sequence. We used the DeepSEA model as our sequence-encoding component.

2.4 Bidirectional GRU architecture

Recurrent neural networks (RNNs) are advanced ANN models that are suited for classification or regression tasks whose inputs or outputs are sequences (Schuster and Paliwal, 1997). RNNs connect all the hidden layers of the neural network, which can effectively address the dependence of the adjacent data in a sequence. RNNs handle the sequential data using a recurrent hidden state, and the activation at each time is dependent on that of the previous time. More formally, given sequential data $x=(x_1, x_2, \cdots, x_t)$, RNNs update their recurrent hidden state h_t by

$$h_t = \begin{cases} 0, & t = 0 \\ \sigma(Wx_t + Uh_{t-1}), & otherwise \end{cases}, \quad (1)$$

where W is the weight matrix of the current hidden layer, U is the weight matrix of the hidden layer in the last time step, and $\sigma(\cdot)$ denotes the nonlinear activation function. Then, a generative RNN outputs a probability distribution that can be decomposed into the following (Chung *et al.*, 2015):

$$p(x_1, \ldots, x_T) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_T|x_1, \ldots, x_{T-1}). \quad (2)$$

Regular RNNs process the input information in a positive time direction only (forward states). BRNNs can be trained using all available information in both the positive time direction (forward states) and the negative time direction (backward states). Since we tend to solve a binary classification task, we use the last element in the equation (2) as our final result, which can be represented by

$$p(x_t|x_1, \ldots, x_{t-1}) = \sigma\left(Wh_t^f + Wh_t^b\right), \quad (3)$$

where f and b denote the forward and backward states, respectively.

In general, backpropagation through time algorithms can be used to effectively train BRNNs. However, difficulties are encountered when analysing relatively long sequential data because the standard BRNN structure can retain only short-term memory due to the vanishing gradient problem. A GRU, which was first designed by Kyunghyun Cho (Bahdanau *et al.*, 2014), was proposed to solve this long-term dependency problem. The GRU architecture works through two gates, reset gate r_t and update gate z_t, and a candidate hidden layer $\tilde{h}_t$. The reset gate controls the information that should be retained in the current output h_t, and the update gate controls the information that should be forgotten from the hidden layer at the previous time h_{t-1}. All operations can be summarized as described in a previous study (Chung *et al.*, 2015):

$$r_t = \sigma(W_r x_t + U_r h_{t-1}), \quad (5)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}), \quad (6)$$

$$\tilde{h}_t = \tanh(Wx_t + r_t \odot Uh_{t-1}), \quad (7)$$

$$h_t = z_t\tilde{h}_{t-1} + (1 - z_t)h_t. \quad (8)$$

Reset gates are typically more effective when applied to short-term memory problems, whereas update gates are important for long sequence data.

2.5 Model design and training of BiRen

To prepare the input for the DeepSEA model, we segmented each VISTA Enhancer element into 200-bp bins, and we extended these 200-bp bins to 1000-bp bins by flanking 400 bp at each side. We then extracted DNA sequence data from the human GRCh37/hg19 reference genome and represented the 1000-bp DNA sequence by a 1000×4 binary matrix with columns corresponding to A, G, C and T. Next, we used DeepSEA to encode each 200-bp bin with a label vector for 919 chromatin features, which represented the possibility of 919 chromatin features for each 200-bp sample. These 919 chromatin features consist of 125 DNase features, 690 transcription factor (TF) features and 104 histone features.

After we collected all 920 features, including the 919 chromatin features and PhastCons conservation scores for each 200-bp sample, we rebuilt each enhancer element with these 200-bp samples according to their original location in the enhancer element. Thus, each enhancer element was represented as a $920 \times N$ matrix, where N is the number of 200-bp units of each enhancer element. Given that the length of each enhancer element was different and the maximum length was 8061 bp (40 200-bp units), we extended this matrix to a 920×40 matrix by padding zeros for each enhancer element. Additionally, we constructed a binary mask vector of the same length, with 1 representing the value unit and 0 representing the zero unit for each enhancer element.

According to the nature of our enhancer sequence set, we selected a GRU-based BRNN. Except for the input and output layers, our BiRen model used $N=40$ bidirectional GRU units that shared the same input and the same structure. Both the forward-GRU layer and the backward-GRU layer had one hidden layer with 140 neurons and were concatenated with one concatenation layer to merge all parameters. All weights were initialized by randomly drawing from uniform distribution, and all bias terms were initially set to 0.

Our GRU-BRNN models were trained using the AdaDelta algorithm (Zeiler, 2012). We set the squared gradient moving average decay factor (*rho*) to 0.95 and set the numerical stability score (*epsilon*) to *1e-7*. Thus, the learning rate was no longer considered, and the parameters were self-adapted, as described in a previous study (Zeiler, 2012).

Our BiRen was implemented using the Lasagne 0.2.0 library (https://github.com/Lasagne/Lasagne).

2.6 Validation of enhancer predictions

To validate the enhancers predicted by BiRen and other existing methods, we calculated the validation rate, which was defined as the percentage of the predicted enhancers that overlapped with enhancer markers, chromatin states associated with enhancers (Hoffman *et al.*, 2012), and HOT (high-occupancy target) regions (Li *et al.*, 2015, 2016) according to the definition used in our recent study (Liu *et al.*, 2015, 2016). The enhancer markers included distal DHSs of 125 human cells and H3K27ac-binding sites of 86 human cells/tissues, which were derived from the ENCODE project (ENCODE Project Consortium, 2012) and the Roadmap Epigenomics Project (Kundaje *et al.*, 2015), respectively. Strong enhancer segmentations across six human cell types were obtained by the unsupervised machine learning technique Segway (Hoffman *et al.*, 2012). HOT regions of 503 human cells/tissues, which were bound by a surprisingly large number of transcription factors, were derived from the ENCODE project (ENCODE Project Consortium, 2012) and the Roadmap Epigenomics Project (Kundaje *et al.*, 2015; Li *et al.*, 2015, 2016). HOT regions play key roles in cell development and differentiation (Li *et al.*, 2016) and in human disease and cancer (Li *et al.*, 2015). TSS annotations were extracted from the GENCODE annotations (V15) (Harrow *et al.*, 2012). Predicted enhancers overlapping with a window of -100 to +100 bp centred at a distal enhancer marker that was greater than 5 kb away from the nearest TSS were classified as 'validated'.

2.7 Performance assessment and comparison

To assess the performance of BiRen and other existing methods, we used receiver operating characteristic curves (ROCs) that used both the POSITIVE enhancers and the combination of POSITIVE and NEGATIVE enhancers as 'gold standard enhancers'. Additionally, the corresponding the area under the curve (AUC) was computed. To further compare the performance across different methods in the genome-wide prediction of enhancers, validation rates of distal DHSs, H3K27ac and HOT regions were also computed according to the definition used in our recent studies (Liu *et al.*, 2015, 2016).

2.8 Determination of the optimal input window for BiRen

To predict genome-wide enhancers in the human and mouse genomes, it was critical to determine the optimal input window of the DNA sequence because the input sequence determines the regulatory properties of the enhancers and thus is important for identifying enhancer elements. Thus, we used the trained BiRen models to predict enhancer elements in the human genome with input windows ranging from 600 bp to 1400 bp with 200-bp steps. We compared the performance assessment of these cases using the validation rates of distal DHSs, H3K27ac, super-enhancers and HOT regions (Supplementary Fig. S2). We found that our BiRen models with 800-bp input DNA sequences achieved superior performance in the genome-wide prediction of enhancer elements in the human genome.

3 Results

3.1 Prediction of developmental enhancers using the DNA sequence alone

Our primary aim in this study was to identify enhancer elements in a mammalian genome using the DNA sequence directly but not DNA sequence features. To this end, we present a deep-learning-based hybrid architecture named BiRen that integrates a CNN and a GRU-BRNN (Fig. 1, see Materials and methods). In our BiRen model, the CNN encodes the original DNA sequence to a label vector for 919 chromatin features using DeepSEA (Zhou and Troyanskaya, 2015), and GRU-BRNN models the probability of predicted enhancers based on the learned 919 features and evolutionary conservation information. Our BiRen approach combines both the sequence encoding and representation power of CNN and the long-term dependency capacity of GRU-BRNN. Thus, BiRen has excellent potential for the *de novo* discovery of regulatory elements, such as enhancers, using the DNA sequence alone.

To train our BiRen model, we constructed an enhancer set using 900 POSITIVE human enhancers from the VISTA Enhancer Browser (Visel *et al.*, 2007) and a non-enhancer set of random genomic loci not annotated as promoters or enhancers (Materials and methods). First, we split each enhancer element into 200-bp bins and encoded each of the 200-bp bins using DeepSEA with a label vector of 919 chromatin features consisting of 125 DNase features, 690 TF features and 104 histone features (Zhou and Troyanskaya, 2015) in addition to evolutionary conservation scores. Thus, each enhancer element was encoded by a $920 \times N$ matrix, where N is the number of 200-bp bins of each element. To keep the different lengths of the POSITIVE human enhancers consistent, we set $N = 40$, which is the largest number of 200-bp bins in the POSITIVE human enhancers.

Second, we used all the 920×40 matrices and the corresponding binary mask matrices as the inputs to GRU-BRNN. Then, we

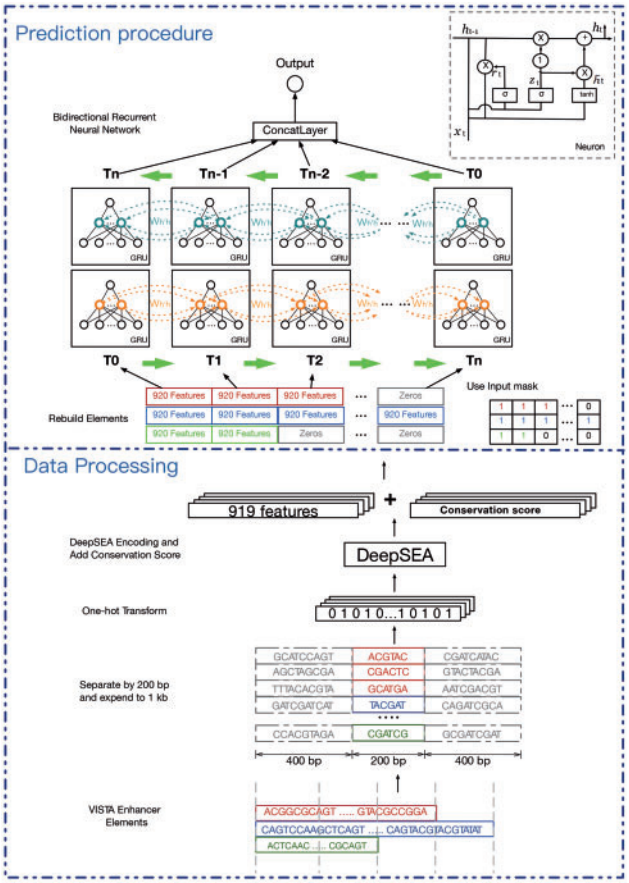

Fig. 1. The workflow of BiRen. Schematic diagram depicting the encoding DNA sequence procedure and prediction procedure of BiRen. DeepSEA encodes each input DNA sequence to a label vector for 919 chromatin features, which are used as inputs for the GRU-BRNN to model the probability of enhancers

trained the GRU-BRNN model and assessed the prediction performance in terms of the ROC curve and its corresponding AUC using 5-fold cross-validation in both the training set and the test set. After a grid search for all optional hyperparameters, we finally obtained the best GRU-BRNN model, which consists of $N = 40$ bidirectional GRU units corresponding to the largest number of 200-bp bins. Each GRU unit is a neural network with 920 inputs, one output, and one hidden layer containing 140 neurons, which has a more sophisticated structure (Fig. 1). All the hidden layers of forward-GRU and backward-GRU in 40 GRU units were concatenated with one concatenation layer to merge all parameters. On average, our BiRen method achieved AUC = 0.945 in the training set and AUC = 0.945 in the test set (Fig. 2A). This result demonstrates the superior performance and outstanding robustness of BiRen in both the training and test sets, suggesting that BiRen has an excellent ability to solve the overfitting problem in enhancer prediction.

3.2 Performance assessment of BiRen

To assess the robustness of the ability of BiRen to overcome the noise of false-positive enhancers in the training set, we combined the 900 POSITIVE elements and 847 NEGATIVE elements in the VISTA Enhancer Browser to construct a new enhancer set. Correspondingly, we generated a non-enhancer set that contained 10 times the number of random genomic loci not annotated as promoters or enhancers. We optimized the BiRen model using the new

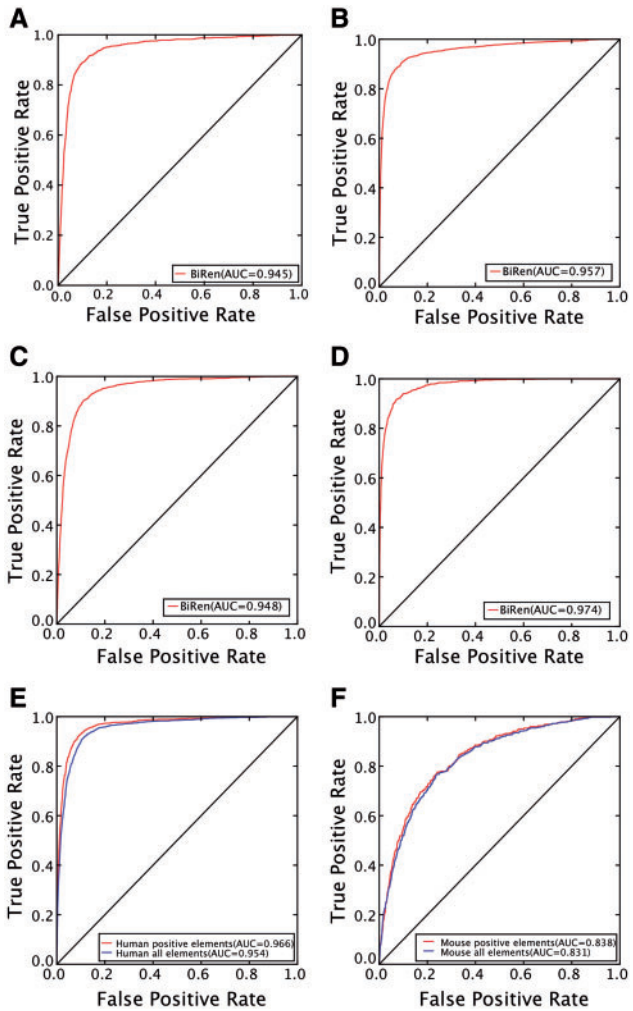

Fig. 2. Performance assessment of BiRen. (**A, B**) ROC curves used to assess the performance of BiRen models trained with human POSITIVE elements (A) and a combination of both human POSITIVE and NEGATIVE elements (B) from the VISTA Enhancer Browser. (**C, D**) Cross-validation of BiRen models trained with two enhancer sets. The BiRen model trained with human POSITIVE elements was used to predict the enhancer set of both human POSITIVE and NEGATIVE elements (C). The BiRen model trained with a combination of both human POSITIVE and NEGATIVE elements was used to predict the enhancer set of human POSITIVE elements (D). (**E, F**) Generalizability assessment of BiRen. (E) The BiRen model trained with human POSITIVE elements was used to predict mouse elements (red: POSITIVE elements; blue: POSITIVE + NEGATIVE elements). (F) The BiRen model trained with mouse POSITIVE elements was used to predict human elements (red: POSITIVE elements; blue: POSITIVE + NEGATIVE elements)

enhancer set and non-enhancer set and assessed the prediction performance in terms of the ROC curve and its corresponding AUC using 5-fold cross-validation in both the training set and the test set (Fig. 2B). On average, we still obtained AUC = 0.944 and AUC = 0.957 in the training set and test set, respectively, values that were similar to those obtained with only POSITIVE elements. Then, we used the BiRen models that were trained using only POSITIVE elements and a new enhancer set to perform cross-validation (Fig. 2C and D). The BiRen model trained using only POSITIVE elements was used to predict a new enhancer set and produced AUC = 0.948 (Fig. 2C), whereas the BiRen model trained using the new enhancer set was used to predict POSITIVE elements and produced AUC = 0.974 (Fig. 2D). These results indicate that the BiRen model provides superior anti-noise performance.

To further assess the generalizability of our BiRen model, we used our optimized BiRen model trained with 900 human POSITIVE elements to predict experimentally validated mouse non-coding elements from the VISTA Enhancer Browser. We collected 322 POSITIVE and 245 NEGATIVE mouse elements from the VISTA Enhancer Browser. We obtained AUC = 0.838 and AUC = 0.831 in the POSITIVE mouse element set and in both POSITIVE and NEGATIVE mouse element sets, respectively (Fig. 2E). Additionally, we generated a non-enhancer set in the mouse genome that contained 3220 random genomic loci not annotated as promoters or enhancers and trained the BiRen model with 322 mouse POSITIVE elements using 5-fold cross-validation. On average, our BiRen method achieved AUC = 0.926 in the training set and AUC = 0.908 in the test set (Supplementary Fig. S1). We used the trained BiRen model to predict human enhancer elements in the same manner (Fig. 2F). We obtained AUC = 0.966 and AUC = 0.954 in the POSITIVE human element set and both POSITIVE and NEGATIVE human element sets, respectively. Together, our results suggest that our BiRen trained with human enhancer elements can be well generalized to predict mouse enhancer fragments and vice versa.

3.3 Validation of predicted enhancers

To validate the enhancers predicted with BiRen on a genome-wide scale, we calculated the validation rate as the percentage of predicted enhancers overlapping distal DHSs across 125 cell lines, H3K27ac in 86 human cell and tissue types, strong enhancer segmentations in six cell types (Hoffman *et al.*, 2012), and HOT regions across 503 human cell and tissue types (Li *et al.*, 2015, 2016) (see Materials and methods). We used 45 036 enhancer elements in the human genome identified by BiRen with an optimal 800-bp input DNA sequence, with which the BiRen model was superior in the genome-wide prediction of enhancers with different sizes of input windows (Supplementary Fig. S2). This result suggests that our hybrid architecture combining CNN- and GRU-based RNNs allows us to scale to accommodate long sequence inputs and to learn sequence dependencies. For distal DHSs, H3K27ac, HOT regions and strong enhancer segmentations, BiRen achieved validation rates of 55.0% ± 0.08, 40.1% ± 0.11, 23.4% ± 0.07 and 27.4% ± 0.13, respectively (Table 1). Distal DHSs, H3K27ac and HOT regions were enriched in enhancer sequences, suggesting that sequences with all these features would be more accurately identified as enhancers. Thus, we computed the percentage of predicted enhancers marked with all three epigenetic features across the 20 common human cell/tissue types, and still obtained a validation rate of 15.3% ± 0.07 for BiRen (Table 1 and Supplementary Table S1). These results demonstrate that BiRen can accurately predict putative enhancers using the DNA sequence alone.

3.4 Performance comparisons of BiRen with existing methods

We compared the performance of BiRen with those of the two existing methods that identify enhancer elements from DNA sequence. Similar to our BiRen model, the DEEP-VISTA model uses enhancers archived in the VISTA Enhancer Browser to train SVM models with 351 attributes derived from the sequences themselves (Kleftogiannis *et al.*, 2015a,b). Lee *et al.* (2011) developed an SVM framework that accurately identifies mammalian enhancers using genomic sequence features of the full set of *k*-mers ($k = 3$–10 bp).

To achieve a fair performance comparison between our method and these two supervised approaches, we applied the trained BiRen

Table 1. Performance comparisons across methods with comparable enhancer predictions

Methods	Number of predictions	DHS	H3K27ac	HotRegion	Enhancer Segmentation	DHS+H3K27ac +HOTregion
BiRen	45036	55.0%±0.08	40.1%±0.11	23.4%±0.07	27.4%±0.13	15.3%±0.07
DEEP VISTA	47100	29.2%±0.08	34.3%±0.10	20.7%±0.06	10.8%±0.10	9.0%±0.06
Lee's SVM	43135	34.1%±0.10	24.2%±0.08	16.9%±0.06	19.8%±0.10	11.2%±0.06

Note: Performance comparisons across methods with comparable enhancer predictions with a threshold defined by the BiRen method.

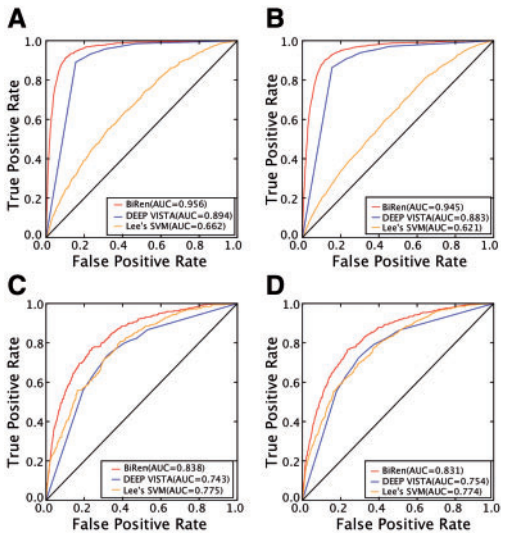

Fig. 3. Performance comparison between BiRen and existing methods. (**A, B**) Performance comparisons across methods with ROC curves and AUCs using human POSITIVE enhancers (A) and human POSITIVE and NEGATIVE enhancers (B) as 'gold standard enhancers'. (**C, D**) Performance comparisons across methods with ROC curves and AUCs using mouse POSITIVE enhancers (C) and mouse POSITIVE and NEGATIVE enhancers (D) as 'gold standard enhancers'

model and the two supervised methods to predict enhancers in the human and mouse genomes based on their optimal parameters. We used the VISTA human elements as 'gold standard enhancers' to assess the performance of these predictions with ROC curves and AUCs (Fig. 3). Our results suggest that BiRen predicted enhancers with greater accuracy; the AUCs were 0.956, 0.894 and 0.662 for BiRen, DEEP-VISTA and Lee's SVM with POSITIVE enhancers as 'gold standard enhancers', respectively (Fig. 3A). We also obtained consistent results when using the combination of POSITIVE and NEGATIVE human enhancers as 'gold standard enhancers' (Fig. 3B). Next, we used BiRen, DEEP-VISTA and Lee's SVM method separately to predict enhancers in the mouse genome (Fig. 3C and D). Our BiRen method also enabled higher-performance prediction using both POSITIVE enhancers (AUC = 0.838) and the combination of POSITIVE and NEGATIVE enhancers (AUC = 0.831) as 'gold standard enhancers' relative to the other two methods.

For the four cases of 'gold standard enhancers', we further compared the selectivity (false-positive rate) across the three methods at sensitivity (true-positive rate) levels of 0.5, 0.8 and 0.95 (Supplementary Table S2). Our BiRen consistently achieved the best selectivity (lowest false discovery rate) compared with the other two supervised methods, suggesting superior performance relative to the other two state-of-the-art enhancer predictors.

Additionally, we performed a comparative analysis across BiRen, DEEP-VISTA and Lee's SVM method in the genome-wide identification of enhancers in the human genome. To exclude bias due to the different numbers of enhancer predictions made by these methods in the comparison of performance, we selected thresholds that yielded comparable numbers of predictions for DEEP-VISTA and Lee's SVM method to perform fair comparisons across these methods. We obtained 47 100 and 43 135 enhancer predictions for DEEP-VISTA and Lee's SVM, respectively, which was comparable to the number of predictions obtained by our BiRen (45 036 predictions) (Table 1). We compared BiRen with DEEP-VISTA and Lee's SVM method in the genome-wide prediction of enhancer elements based on validation rates of distal DHSs, H3K27ac, HOT regions and strong enhancer segmentations. The comparative analysis across these methods indicated that BiRen consistently performed better than all the other methods (Table 1). Even based on the percentage of predicted enhancers marked with distal DHSs, H3K27ac and HOT regions, BiRen (15.3%) was ranked first, followed by SVM (11.2%) and DEEP (9.0%), across the 20 common human cells/tissues of all three epigenetic features (Tables 1 and S1). This result suggested that the accuracy rate of BiRen was higher than those of the other two methods. Furthermore, we also used the optimal thresholds of both DEEP-VISTA and Lee's SVM method to predict enhancer elements at the genome-wide scale, and we selected thresholds that yielded comparable numbers of predictions for the other two methods to separately perform fair comparisons across these methods (Supplementary Tables S1 and S3–S4). In both cases, our BiRen manifested superior performance consistency relative to the two supervised methods. Taken together, these results demonstrate that our BiRen method exhibits superior performance relative to the existing methods of enhancer prediction.

4 Discussion

In this study, we proposed BiRen to precisely identify enhancer elements on a genome-wide level using the DNA sequence alone. To resolve the major obstacle of the lack of large, sufficiently comprehensive and experimentally validated enhancers in the enhancer identification problem, BiRen adopts a deep-learning-based hybrid structure that is trained with limited experimentally validated non-coding elements derived from the VISTA Enhancer Browser (Visel *et al.*, 2007) that have gene enhancer activity, as assessed in transgenic mice. The BiRen hybrid model integrates the sequence encoding and representation power of CNN and the superior capacity for handling the long-term dependency of the long DNA sequence of GRU-BRNN to successfully resolve the challenge of the identification of enhancers using the DNA sequence alone. Subsequent performance assessments demonstrated that our BiRen approach achieves substantial improvements in identification accuracy, robustness in overcoming noise data, and generalization to other species for enhancer predictions relative to two state-of-the-art methods based on sequence characteristics, such as motifs or *k*-mers. Our results suggest that our BiRen can help decipher the transcriptional regulatory code encoded in the four-letter 'alphabet' of enhancer sequences in human and mouse genomes.

Although our BiRen demonstrates superior learning power from limited experimentally validated enhancers, its predictive capability remains weaker than the enhancer predictors that are well trained with cell-type-/tissue-specific enhancer markers or their combinations. This weakness arises because enhancer predictors based on epigenetic profiles, DHSs and TFBSs can take advantage of a large variety of enhancers as training and test sets across diverse cell types/tissues from different species. With the further knowledge gained by studying enhancers on a genome-wide level and the enriching and growing number of functionally validated enhancers obtained by enhancer testing and screening methods, BiRen will enable the systematic and comprehensive identification of enhancers and will provide key insights into the regulatory functions that are encoded in enhancer sequences.

Acknowledgements

We wish to thank the ENCODE Project Consortium for making their data publicly available. The authors would also like to thank the anonymous reviewers for their constructive comments, which contributed to an improved presentation of our study.

Funding

This work was supported by grants from the Major Research Plan of the National Natural Science Foundation of China (No. U1435222), the Program of International S&T Cooperation (No. 2014DFB30020) and the National High Technology Research and Development Program of China (No. 2015AA020108).

Conflict of Interest: none declared.

References

Bahdanau,D. *et al.* (2014) Neural machine translation by jointly learning to align and translate. *Comput. Res. Reposit.*, abs/1409.0473.

Bulger,M. and Groudine,M. (2010) Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, **339**, 250–257.

Bulger,M. and Groudine,M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327–339.

Calo,E. and Wysocka,J. (2013) Modification of enhancer chromatin: what, how, and why?. *Mol. Cell*, **49**, 825–837.

Chung,J. *et al.* (2015) Gated feedback recurrent neural networks. In: *32nd International Conference on Machine Learning*. ICML, Lille, France. pp. 2067–2075.

Cornu,T.L. and Milner,B. (2015) Voicing classification of visual speech using convolutional neural networks. In: *1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*. ISCA, Vienna, Austria. pp. 103–108.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Fernandez,M. and Miranda-Saavedra,D. (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res.*, **40**, e77.

Firpi,H.A. *et al.* (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics (Oxford, England)*, **26**, 1579–1586.

Goldman,M. *et al.* (2015) The UCSC cancer genomics browser: update 2015. *Nucleic Acids Res.*, **43**, D812–D817.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.

Hnisz,D. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Kleftogiannis,D. *et al.* (2015a) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.*, **43**, e6.

Kleftogiannis,D. *et al.* (2015b) Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinf.*, **17**, 967–979.

Krizhevsky,A. *et al.* (eds.) (2012) *Advances in Neural Information Processing System*. NIPS Proceedings, Lake Tahoe. pp. 1097–1105.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Lawrence,S. *et al.* (1997) Face recognition: a convolutional neural-network approach. *IEEE Trans. Neural Netw.*, **8**, 98–113.

Lee,D. *et al.* (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, **21**, 2167–2180.

Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Li,H. *et al.* (2015) Functional annotation of HOT regions in the human genome: implications for human disease and cancer. *Sci. Rep.*, **5**, 11633.

Li,H. *et al.* (2016) Genome-wide identification and characterisation of HOT regions in the human genome. *BMC Genomics*, **17**, 733.

Liu,F. *et al.* (2016) PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.*, **6**, 28517.

Liu,F. *et al.* (2015) De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics (Oxford, England)*, **32**, 641–649.

Loven,J. *et al.* (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320–334.

Lu,Y. *et al.* (2015) DELTA: a distal enhancer locating tool based on AdaBoost algorithm and shape features of chromatin modifications. *PloS One*, **10**, e0130622.

Meng,F.L. *et al.* (2014) Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell*, **159**, 1538–1548.

Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.

Rajagopal,N. *et al.* (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, **9**, e1002968.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process*, **45**, 2673–2681.

Shlyueva,D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Stathopoulos,A. and Levine,M. (2005) Genomic regulatory networks and animal development. *Dev. Cell*, **9**, 449–462.

Visel,A. *et al.* (2007) VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.

Visel,A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199–205.

Whyte,W.A. *et al.* (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.

Zeiler,M.D. (2012) ADADELTA: an adaptive learning rate method. *Comput. Res. Reposit.*, abs/1212.5701.

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