Briefings in Bioinformatics, 22(4), 2021, 1–8

https://doi.org/10.1093/bib/bbaa356
Problem Solving Protocol

DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops

Fu-Ying Dao, Hao Lv, Dan Zhang, Zi-Mei Zhang, Li Liu and Hao Lin

Corresponding authors: Hao Lin, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China. Tel: +86-13678168394; E-mail: hlin@uestc.edu.cn; Li Liu, Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China. Tel: +86-15661178088; E-mail: liliu2010imu@163.com

Abstract

The protein Yin Yang 1 (YY1) could form dimers that facilitate the interaction between active enhancers and promoter-proximal elements. YY1-mediated enhancer–promoter interaction is the general feature of mammalian gene control. Recently, some computational methods have been developed to characterize the interactions between DNA elements by elucidating important features of chromatin folding; however, no computational methods have been developed for identifying the YY1-mediated chromatin loops. In this study, we developed a deep learning algorithm named DeepYY1 based on word2vec to determine whether a pair of YY1 motifs would form a loop. The proposed models showed a high prediction performance (AUCs≥0.93) on both training datasets and testing datasets in different cell types, demonstrating that DeepYY1 has an excellent performance in the identification of the YY1-mediated chromatin loops. Our study also suggested that sequences play an important role in the formation of YY1-mediated chromatin loops. Furthermore, we briefly discussed the distribution of the replication origin site in the loops. Finally, a user-friendly web server was established, and it can be freely accessed at http://lin-group.cn/server/DeepYY1.

Key words: YY1-mediated chromatin loops; deep learning; word2vec; the distribution of the replication origin site; enhancer–promoter interaction

Introduction

Higher order chromatin structure is critically important for basic biological processes, such as cell differentiation, transcriptional regulation, genome replication and DNA repair [1]. The spatial structure of genome organization is hierarchical, and its most basic structural unit is called the topologically associating domain (TAD) [2]. Specific DNA interactions between the proximal promoters and distal regulatory elements form a chromatin loop within these relatively stable chromatin domains [3]. Large DNA loops encompassing genes and their regulatory elements depend on the chromatin architectural protein CCCTC-binding factors (CTCF) interactions. CTCF and the associated cohesin complex play a central role in insulator function and higher-order chromatin organization of mammalian genomes as shown in Figure 1A. However, most enhancer–promoter interactions do not employ structural protein CTCF. The protein Yin Yang 1 (YY1) [4, 5] has been reported to bind hypo-methylated DNA sequences form homodimers, and it can contribute to enhancer–promoter

Fu-Ying Dao is a PhD candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. Her research interests are bioinformatics, machine learning and DNA replication regulation.
Hao Lv is a PhD candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. His research interests are bioinformatics, machine learning and DNA and RNA modification.
Dan Zhang is a MS candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. Her research interests are bioinformatics and machine learning.
Zi-Mei Zhang is a MS candidate of the Center for Informational Biology at the University of Electronic Science and Technology of China. Her research interests are bioinformatics.
Li Liu is a lecturer of the Laboratory of Theoretical Biophysics at the Inner Mongolia University. His research field is bioinformatics.
Hao Lin is a professor of the Center for Informational Biology at the University of Electronic Science and Technology of China. His research is in the areas of bioinformatics and system biology.
Submitted: 21 September 2020; **Received (in revised form):** 19 October 2020

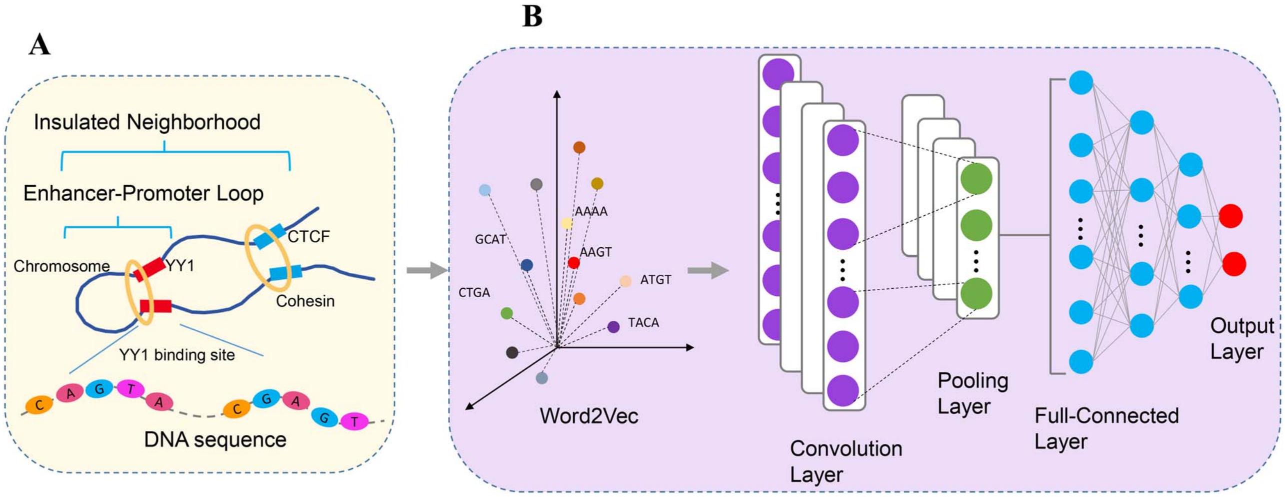


Figure 1. The outline of DeepYY1. (**A**) YY1 is a structural regulator of enhancer–promoter interactions and facilitates gene expression. (**B**) Visualization of the detailed architecture of DeepYY1.

interactions in a manner analogous to DNA looping mediated by CTCF [6] (Figure 1**A**). The majority of YY1 preferentially occupies interacting enhancers and promoters, and the loop structure mediated by YY1 is related to gene activation and repression and to gene dysregulation in cancer.

Precise identification of interactions between regulatory elements is critically important not only for elucidating transcriptional regulation but also for understanding the mechanisms underlying complex human diseases [7]. Dissecting the factors underlying the formation and maintenance of genome structure requires high-throughput sequencing. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA–PET) [8] and High-through chromosome conformation capture (Hi-C) [9, 10] techniques can capture the chromosome interaction regions, and massive sets of experimental data have been accumulated. In addition, recent advances in deep learning methods have provided novel alternative approaches for chromatin interaction identification [7, 11–13]. Artificial intelligence can complement wet experiments for characterizing chromatin interactions [14, 15].

To date, many computational methods have been developed to detect chromatin interactions at the level of regulatory elements [11–13, 16–22]. Most of the aforementioned studies are based on the genomic signature information and CTCF binding site information in large DNA loops. Weintraub *et al.* [6]. found YY1 is a structural regulator of enhancer-promoter loops in a manner analogous to the DNA interactions mediated by CTCF. YY1 preferentially occupies interacting enhancers and promoters, whereas CTCF preferentially occupies sites distal from these regulatory elements that tend to form larger loops and participate in insulation (Figure 1**A**). Therefore, the use of YY1 binding site information is important for the identification of small DNA loops.

Here, we developed a deep learning algorithm based on word embedding called DeepYY1 to predict if a pair of YY1 motifs would form a loop using the YY1 binding site sequence features. The DeepYY1 method has a high predictive ability with high AUC values. We show that our method can produce models with a high discriminating ability and enhance our understanding of the factors controlling YY1-mediated chromatin loops. Generally, we emphasize that the DNA sequence contains important information for complex chromatin architectures.

Materials and methods

Data collection and preprocessing

Mumbach *et al.* [23] presented a protein-centric chromatin conformation method named HiChIP, which combined the advantages of Hi-C and ChIA–PET to obtain higher-resolution three-dimensional (3D) chromatin structural information with smaller datasets. Weintraub *et al.* [6] conducted YY1 HiChIP experiments in two different human cell types, including K562 and HCT116. They found that YY1 generally occupies sites involved in enhancer–promoter interactions in mammalian cells. Therefore, we downloaded YY1 HiChIP and YY1 ChIP-seq data of K562 and HCT116 cell types, respectively.

We defined the positive sample as the HiChIP chromatin loop (confidence probability $\geq$ 0.9), with either side of the paired regions having a unique ChIP-seq YY1 peak. We defined the negative sample as the HiChIP loop with confidence probability = 0. For each pair in both the positive and negative samples, a sequence of 506 bp with YY1 motif as the center was extracted. The YY1 motif PWM was obtained from JASPAR (ID: MA0095.1) [24, 25].

To objectively evaluate the proposed models, we separated the benchmark dataset into two parts: the independent dataset (testing dataset) and the training dataset, according to the ratio of 3:7 [26]. The training dataset was used to train the model by 10-fold cross-validate strategy, whereas the independent dataset was used to test the corresponding obtained model. Details on these benchmark datasets have been provided in Supplementary Table S1 available online at https://academic.oup.com/bib.

Feature description

Formulating DNA sequences with an effective mathematical expression is a key step to build classification model [27, 28], as well as choosing discriminative vector features facilitates to generate model with a high-generalization ability.

Word2vec

The popular word2vec [29] method is a two-layer neural network model, which uses distributed representation to reduce word vectors from high-dimensional to low-dimensional space to solve the dimensional disasters often encountered in deep learning. In natural language processing (NLP) [30], the quality of a model depends largely on the effect of word vectors. Word2vec converts independent words into dense vectors. In such a model, words that appear in the same contexts share semantic meaning; that is, words are embedded in a continuous vector space where 'semantically similar' words have closer vectors.

The 506 bp sequence, which was obtained by the above method, was scanned by a k-sized sliding window with a step size of 1 bp to convert multiple subsequences of fixed length k [31]. We considered subsequences of fixed length k as DNA 'words'. The lengths of all of the YY1 motif sequences in both the positive samples and negative samples were then amplified 250 bp upstream and downstream, respectively. The collection of all possible 'words' in DNA sequences of length 506 bp was defined as the vocabulary.

Here, we utilized word2vec to extract DNA sequence features in a vector space based on a continuous bag-of-words (CBOW) [32] model that involved using the context of each word as the input and attempting to predict the word corresponding to the context. There are three layers in the model: the input layer, hidden layer and output layer. W and W' are the shared input weight matrix and output weight matrix, respectively. The input layer of the model is a word vector. Since the CBOW model obtains the target word through n predictions before and after the target word as shown in Equation (1), the target function of the model can be easily obtained as follows:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\log P\left(w_t|w_{t-n},\cdots,w_{t-1},w_{t+1},\cdots,w_{t+n}\right) \tag{1}$$

where w_t is the target word, and w_{t-n}, ..., w_{t+n} represent the context words. Because the hidden layer does not involve any nonlinear transformation, it can be regarded as a softmax layer so that $P(w_t|w_{t-n},\cdots,w_{t+n})$ can be defined by

$$P\left(w_t|w_{t-n},\cdots,w_{t+n}\right) = \frac{\exp\left(W_t^{T}h_t\right)}{\sum_{k=1}^{v}\exp\left(W_k^{T}h_t\right)} \tag{2}$$

where h_t is the value of the input word vector mapped to the hidden layer vector. The input vector is first subjected to matrix operation with the matrix W, and the average value of all input vectors after matrix operation is obtained to obtain h_t; W' is the hidden layer to the output weight matrix between layers.

We used the gensim software package [33] to train the DNA sequence lexicalization vectors, which is a Python-based NLP library to extract semantic information from the documents. When training the lexicalization vectors by the CBOW model, the sliding window size of the training parameters is set to 20, and the dimension of the target vector is 100. Because each lexicalization vector has important attributes, the minimum filtering frequency of the model was set to 1; that is, no words were filtered, and the number of iterations was 50.

After training the word2vec model, we obtained a hash table with its keys as DNA words and its values as vectors. We then encoded DNA sequences by the embedding vectors for every DNA word in the DNA sentence and taking the average of the vectors as the vector for the DNA sequence. In this study, we set the value of 6 as the word length.

Convolutional neural networks

Because deep learning [34] can efficiently capture local contextual features from unstructured data, the convolutional neural network (CNN) technique [35] was used in this study to construct the YY1-mediated chromatin loops prediction model. As illustrated in Figure 1**B**, the basic architecture of the CNN consisted of six distinct layers: one convolutional layer, one pooling layer, three fully connected layers and one softmax layer. First, the encoding array generated by word2vec connects directly with the convolution layer which scanned the encoding array and resulted in a feature vector. Second, we used a max pooling layer and selected the maximum neuron output value of the local patterns to be the output. Third, the output for each layer is generated by fully connected layers based on this vector. Finally, the output vector of the fully connected layer is used as the input of the softmax layer, which gives the corresponding classification probability to a query sequence.

This newly constructed YY1-mediated chromatin loops prediction model was implemented with Python programming language and the TensorFlow library [36]. Detailed information on the parameters of this CNN strategy is available at http://lin-group.cn/server/DeepYY1/download.html.

Results and discussion

Compare word2vec with other feature extraction methods

According to the data and features described in the Materials and methods, we developed a machine learning approach to predict whether a pair of YY1 motifs can form a chromatin loop. Figure 1 illustrates the workflow of our algorithm, named DeepYY1. For each training dataset of different cell types, we first extracted feature vectors by the word2vec algorithm to generate a feature set, in which we took the value of k as 6. Next, the prediction models were trained using CNN in a 10-fold cross-validation test [37]. In this study, we compared word2vec with other feature extraction methods, such as nucleic acid composition (NAC), dinucleotide composition (DNC), trinucleotide composition (TNC), k-mer, electron–ion interaction pseudopotential (EIIP) value and nature vector (NV) [38, 39]. Here, we used AUC values [40] as the evaluation index because they can describe the overall performance of the model. The receiver operating characteristic (ROC) curves in Supplementary Figure S1 available online at https://academic.oup.com/bib showed the prediction performances of different feature extraction methods based on CNN with 10-fold cross-validation. It is obvious that the word2vec always produced superior results (AUCs ≥0.93) in both of the cell types. In addition, we noticed that k-tuple nucleotide composition (NAC, DNC, TNC and 4-mer) are also good features. For example, the AUC value of 4-mer was better than that of word2vec in the HCT116 cell. Therefore, we further analyzed the models generated by word2vec and the traditional k-mer method based on same k values.

For a more intuitive comparison, the AUC values and program running time generated by the two algorithms were recorded in Figure 2 to further demonstrate word2vec's ability to encode DNA sequences. From the comprehensive view of the two cell lines, we can draw the following conclusions. First, in k562 and HCT116 cell lines, the word2vec method can lead to the stable performance of models based on different k-mer words, as all AUC values were greater than 0.93 (lines of dashes). However, the performance of the models generated by the traditional k-mer method fluctuated sharply as the feature dimension increases;

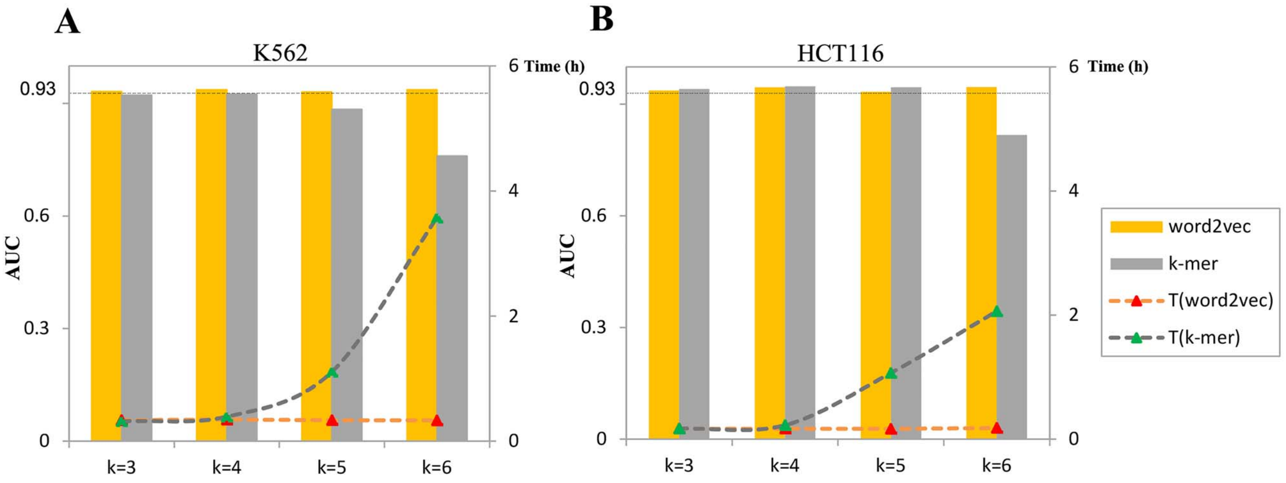


Figure 2. Comparison of model performance and running time based on word2vec and traditional *k*-mer methods in K562 (**A**) and HCT116 (**B**).

models had the lowest performance when $k = 6$. Thus, word2vec can capture more information on short-range sequences while ensuring that the performance of the model is sufficient to maintain its stability. However, model performance decreases under the traditional *k*-mer method when longer short-range sequence information is extracted, resulting in a non-robust model. In addition, these results indicate the effectiveness of the word2vec method for extracting motifs with long segments. Second, according to the running time curves in Figure 2**A** and **B**, when $k \leq 4$, the calculation time of the two feature extraction methods was similar. However, when $k > 4$, the running time increased sharply for the model based on the traditional *k*-mer method, while word2vec was computationally efficient and saved computing resources.

Therefore, word2vec is essentially a dimensionality reduction operation, and the generated dimension (consistent with the number of hidden layer nodes) is generally much smaller than the total number of words. Whereas the vector dimension produced by the traditional *k*-mer method increases exponentially with the increase of *k*, which greatly consumes computational resources. What's more, word2vec can better express the similarity and analogy relationship between different words and show the relevant information between words on the semantic level to better dig out the connections between the words. It is better than traditional *k*-mer method for representing sequence information. Meanwhile, this finding also demonstrates the importance of considering the computational efficiency of feature coding methods to ensure high performance of the model. Taken together, these results suggest that the sequence-based features obtained by word2vec are informative for predicting YY1 chromatin loops.

Analysis of the features extracted by word2vec

We used the CNN with 10-fold cross-validation to train sequence-based feature sets encoded by word2vec, which was named DeepYY1. DeepYY1 was superior for the construction of a robust classification model based on the above analysis of model performance. However, how the feature set from word2vec contributes to improving model performance remains unclear. To answer this question, we visualized the distribution of the samples in the vector space based on the features obtained from word2vec. *t*-distributed stochastic neighbor embedding (*t*-SNE) assigns each data point a location in a two-dimensional (2D) or 3D map and is particularly well suited for the visualization of high-dimensional datasets [41].

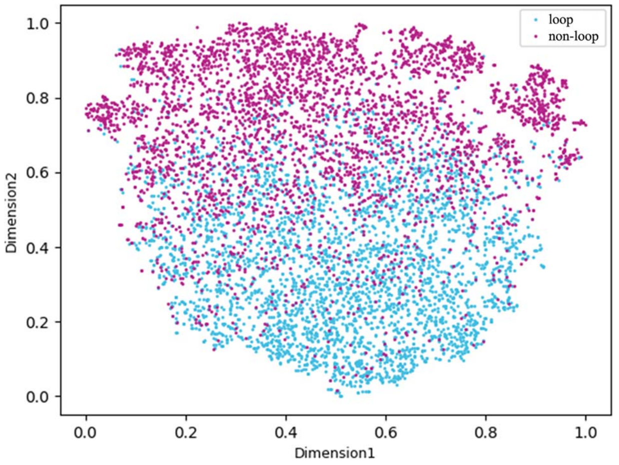


Figure 3. Visualization using *t*-SNE based on features from word2vec that were used in DeepYY1.

We applied *t*-SNE to reduce the feature space to a 2D space and plotted a 2D plane. As shown in Figure 3, the *x*-axis and *y*-axis form the 2D plane, and each point represents a sample, with its color representing the label (either forming a loop or non-loop). We found that most of the positive and negative samples in the feature space were distributed in two clear clusters, suggesting that word2vec can learn the representative features of the loop and non-loop to contribute to the construction of an efficient vector space for loop and non-loop classification. This also explained the result of feature selection in Supplementary Figure S2 available online at https://academic.oup.com/bib, in which the model performance improves with the increase of the feature number in the incremental feature selection (IFS) procedure based on *F*-score [42]. Therefore, such nearly perfect features do not require further feature selecting.

Finally, many machine learning algorithms, including artificial neural network (ANN) [43], *K*-nearest neighbor algorithm (*K*-NN) [44], Naive Bayes (NB) [45], random forest (RF) [46] and AdaBoost [47] were used to compare with CNN to find the best model based on same features produced by word2vec. As shown

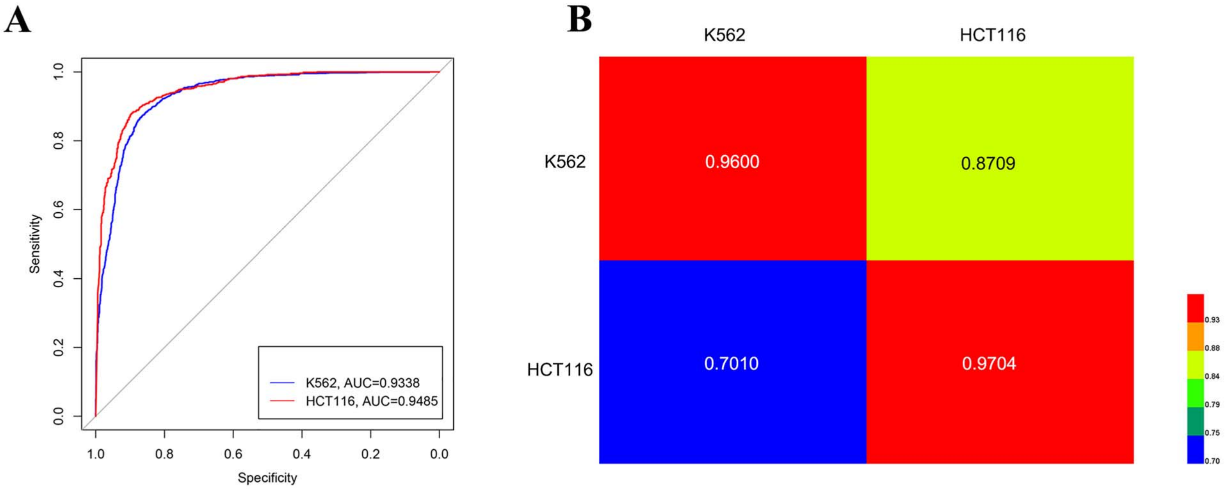


Figure 4. Independent datasets validation and cross-cell validation were used to analyze the robustness and reliability of propose models. (**A**) ROC curves of independent datasets for different cell types; AUC values are 0.9338 and 0.9485 for K562 and HCT116, respectively. (B) The heat map showing the values of AUC in cross-cell types prediction. Once a cell-specific model was established on its own training dataset in rows, it was validated on the data from the same cell as well as with the independent data from the other datasets in columns.

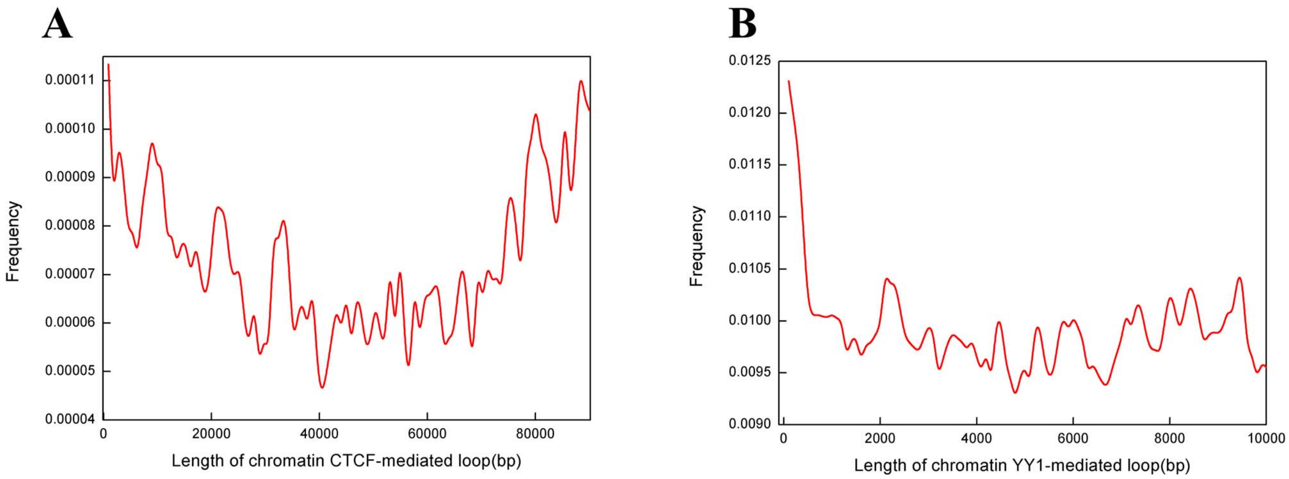


Figure 5. The distribution of ORIs in chromatin loops. (**A**) ORIs tended to be located at the base of the CTCF-mediated chromatin loops. (**B**) ORIs were enriched in the initial portion of the YY1-mediated chromatin loops.

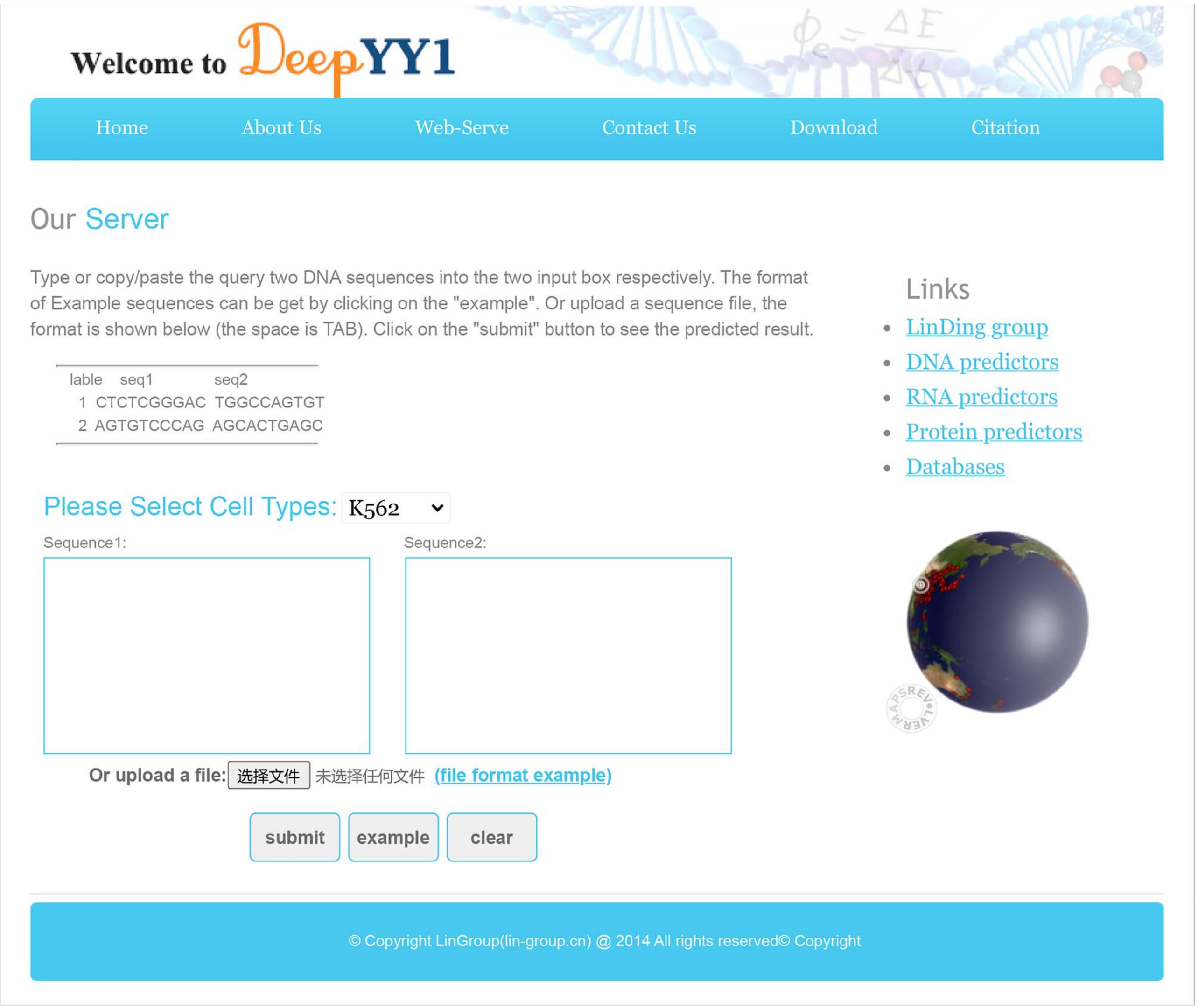
in Supplementary Figure S3 available online at https://academic.oup.com/bib, CNN is superior to other classification algorithms both in two cell types. Hence, the optimal classification models were obtained by a combination of word2vec and CNN.

The robustness and reliability analysis

To further investigate the robustness and stability of the proposed models, we designed following experiments.

Independent datasets have often been used to further test the robustness and stability of the proposed classification models [48]. Minimal over-fitting on the independent dataset indicates that the model has a strong generalization ability. We established independent datasets for each cell type as shown in Supplementary Table S1 available online at https://academic.oup.com/bib. The corresponding results for the different independent datasets are shown in Figure 4A using ROC curves [49]. The AUC values of the independent test in K562 and HCT116 were 0.9338 and 0.9485, respectively. These results indicate that our classification models based on DeepYY1 were sufficient for identifying the YY1-mediated chromatin loops for unknown sequences.

In addition, we conducted a cross-cell validation based on the model described above. Therefore, we can see whether a model trained with the data from one cell type could recognize the interaction between two sequences in other cell types. The knowledge of transfer information [50] was used to design experiment to study the relationships of interacting cell types. The two cell-specific models were first constructed by training datasets from two different cell types, respectively. Subsequently, for each model, the two cells' training datasets were regarded as independent testing datasets to evaluate the performance of the models. A heat map was drawn according to the obtained AUCs in Figure 4B. The models in the rows were tested on the other datasets in columns. It can be seen that the model based on its own dataset always achieves the best accuracy (AUCs > 0.96) but not good results for other cell type dataset, which indicated that the specificity of YY1 binding site sequence among the different



Our Server

Type or copy/paste the query two DNA sequences into the two input box respectively. The format of Example sequences can be get by clicking on the "example". Or upload a sequence file, the format is shown below (the space is TAB). Click on the "submit" button to see the predicted result.

lable	seq1	seq2
1	CTCTCGGGAC	TGGCCAGTGT
2	AGTGTCCCAG	AGCACTGAGC

Please Select Cell Types: K562

Sequence1:

Sequence2:

Or upload a file: 选择文件 未选择任何文件 (file format example)

Links

- LinDing group
- DNA predictors
- RNA predictors
- Protein predictors
- Databases

© Copyright LinGroup(lin-group.cn) @ 2014 All rights reserved© Copyright

Figure 6. A semi-screenshot to show the page of the DeepYY1 web server. Its website address is available at http://lin-group.cn/server/DeepYY1/.

cell types is obvious. Therefore, we cannot use model of one cell type to identify the YY1-mediated chromatin loop in another cell types.

Distribution of ORIs on chromatin loops

Much recent work has focused on studying the mechanism of DNA replication in the spatial organization of the genome [51–53]. We are highly familiar with research in the fields relating to DNA replication and are interested in the relationship between the DNA replication sites (ORIs) and chromatin loops in the 3D genome. To address this question, we examined the distribution of ORIs on loops from the K562 cell line. ORIs data were downloaded from the database DeOri [54], which is one of the most popular and well-known databases for ORIs in eukaryotic genomes. For a more comprehensive comparison, we also obtained CTCF-mediated chromatin loops data of the K562 cell line from [55]. We have drawn the distribution curves of ORIs both in chromatin CTCF-mediated loops and YY1-mediated loops in Figure 5.

ORIs were mainly concentrated at the two ends of the CTCF-mediated loops; that is, ORIs were located at the bases of the CTCF-mediated loops (Figure 5**A**). This result indicates that multiple ORIs in a single replication domain may share DNA replication activation factors, facilitating synchronous activation [51, 56]. However, a similar distribution was not observed in Figure 5**B** in which the ORIs were enriched in the initial portions of the YY1-mediated loops. These findings show that the distributions of ORIs between the CTCF-mediated loop and YY1-mediated loop differed; nevertheless, the specific mechanism driving these differences requires additional study.

Web server guide

According to proposed models, we built a user-friendly online web server, called DeepYY1, to identify whether a pair of DNA sequences would form a chromatin loop. The web server can is freely available at http://lin-group.cn/server/DeepYY1. When users click the ‘Web-Server’ button, the page will jump to the interface as shown in Figure 6. According to the instructions, users can choose two ways to upload query sequences. The first is to paste the DNA sequences into the two input boxes. It is worth noting that the pasted sequences are not in FASTA format, but fragments of DNA sequences. Users could click on the 'example' button to check the right format of sequence. The second is to upload sequences file directly. For the file format, please refer to the instructions or click link to view. Subsequently, click on the 'submit' button to find the predicted result. Here, we recommend users to use the second way to upload sequences, because the first way can only query one pair of DNA sequences interaction at a time, and the second way can complete the identification of multiple pairs of sequences interactions.

Conclusion

In this study, a CNN-based model, named DeepYY1, was developed to predict YY1-mediate chromatin loops with the sequence-based features captured by the word2vec model. The results of both the 10-fold cross-validation and independent testing demonstrated that word2vec features alone can predict loop-forming YY1 motif pairs with high performance. Overall, we believe that the DeepYY1 method is an important step for improving our understanding of the principles governing YY1-mediated chromatin loops. We hope that this work could be used in future studies focused on decoding the information encoded in the genome sequences responsible for determining complex chromatin architectures. In addition, the mechanism of DNA replication in 3D genomes requires additional study. We also hope that the results of this study motivate future work in this field.

Key Points

- We provide the first study to identify the YY1-mediated chromatin loop with the sequence-based features captured by the word2vec.
- The CNN-based model, named DeepYY1, which combined word2vec and CNNs, provides an algorithm combination perspective.
- Excellent performance of DeepYY1 was testified by 10-fold cross-validation and independent testing.
- The relevant research on the mechanism of DNA replication in 3D genomes was provided.

Author's Contribution

F.-Y.D. contributed to the methodology, coding and writing during the original draft preparation. H.Lv contributed to the data curation and writing. D.Z. contributed to the investigation and coding. Z.-M.Z. contributed to the investigation and methodology. L.L. contributed to the validation and conceptualization. H.Lin contributed to the conceptualization, writing–Reviewing and editing.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work has been supported by the National Nature Scientific Foundation of China (61772119, 61961031) and the Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDJQ0012).

Declaration of interests

The authors declare that they have no competing interests.

Data Availability

We provide the benchmark datasets and Python source code of DeepYY1, which are freely available at http://lin-group.cn/server/DeepYY1/download.html.

References

1. Wang Q, Sun Q, Czajkowsky DM, *et al.* Sub-kb hi-C in *D. melanogaster* reveals conserved characteristics of TADs between insect and mammalian cells. *Nat Commun* 2018;**9**:188.
2. Dixon JR, Selvaraj S, Yue F, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.
3. Guo Y, Xu Q, Canzio D, *et al.* CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 2015;**162**:900–10.
4. Kaufhold S, Garban H, Bonavida B. Yin Yang 1 is associated with cancer stem cell transcription factors (SOX2, OCT4, BMI1) and clinical implication. *J Exp Clin Cancer Res* 2016;**35**:84.
5. Antonio-Andres G, Rangel-Santiago J, Tirado-Rodriguez B, *et al.* Role of Yin Yang-1 (YY1) in the transcription regulation of the multi-drug resistance (*MDR1*) gene. *Leuk Lymphoma* 2018;**59**:2628–38.
6. Weintraub AS, Li CH, Zamudio AV, *et al.* YY1 is a structural regulator of enhancer-promoter loops. *Cell* 2017;**171**:1573–1588.e28.
7. Zhu X, Li H-D, Guo L, *et al.* Analysis of single-cell RNA-seq data by clustering approaches. *Curr Bioinform* 2019;**14**:314–22.
8. Li X, Luo OJ, Wang P, *et al.* Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat Protoc* 2017;**12**:899–915.
9. Capurso D, Tang Z, Ruan Y. Methods for comparative ChIA-PET and Hi-C data analysis. *Methods* 2020;**170**:69–74.
10. Belton JM, McCord RP, Gibcus JH, *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 2012;**58**:268–76.
11. Gan M, Li W, Jiang R. EnContact: predicting enhancer-enhancer contacts using sequence-based deep learning model. *PeerJ* 2019;**7**:e7657.
12. Schwessinger R, Gosden M, Downes D, *et al.* DeepC: predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *bioRxiv* 2019. doi: 10.1101/724005.
13. Singh S, Yang Y, Póczos B, *et al.* Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant Biol* 2019;**7**:122–37.
14. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;**2**:719–31.
15. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;**69S**:S36–40.
16. Zhang R, Wang Y, Yang Y, *et al.* Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 2018;**34**:i133–41.
17. Matthews BJ, Waxman DJ. Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *Elife* 2018;**7**:e34077.

18. Kai Y, Andricovich J, Zeng Z, *et al*. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat Commun* 2018;**9**:4221.
19. Zhu Y, Chen Z, Zhang K, *et al*. Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* 2016;**7**: 10812.
20. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96.
21. Al Bkhetan Z, Plewczynski D. Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. *Sci Rep* 2018;**8**:5217.
22. Yang Y, Zhang R, Singh S, *et al*. Exploiting sequence-based features for predicting enhancer-promoter interactions. *Bioinformatics* 2017;**33**:i252–60.
23. Mumbach MR, Rubin AJ, Flynn RA, *et al*. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016;**13**:919–22.
24. Khan A, Fornes O, Stigliani 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**:D260–6.
25. Mathelier A, Fornes O, Arenillas DJ, *et al*. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016;**44**:D110–5.
26. Lv H, Dao FY, Zhang D, *et al*. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 2020;**23**:100991.
27. Yang W, Zhu XJ, Huang J, *et al*. A brief survey of machine learning methods in protein sub-Golgi localization. *Curr Bioinform* 2019;**14**:234–40.
28. Zhang J, Liu B. A review on the recent developments of sequence-based protein feature extraction methods. *Curr Bioinform* 2019;**14**:190–9.
29. Le Q, Mikolov T. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning—Volume 32, ICML 2014*. pp. II–1188–96. JMLR.org, Beijing, China.
30. Tsuruoka Y. Deep learning and natural language processing. *Brain Nerve* 2019;**71**:45–55.
31. Dao FY, Lv H, Zulfiqar H, *et al*. A computational platform to identify origins of replication sites in eukaryotes. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa017.
32. Zeng FX, Ji YF, Levine MD. Contextual bag-of-words for robust visual tracking. *IEEE Trans Image Process* 2018;**27**:1433–47.
33. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010*. University of Malta, Valletta, Malta.
34. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;**61**:85–117.
35. Valueva MV, Nagornov NN, Lyakhov PA, *et al*. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Math Comput Simul* 2020;**177**:232–43.
36. Rampasek L, Goldenberg A. Tensor flow: Biology's gateway to deep learning? *Cell Syst* 2016;**2**:12–4.
37. Allen DM. The relationship between variable selection and data agumentation and a method for prediction. *Dent Tech* 1974;**16**:125–7.
38. Lv H, Dao FY, Guan ZX, *et al*. iDNA6mA-Rice: a computational tool for detecting N6-Methyladenine sites in rice. *Front Genet* 2019;**10**:793.
39. Chen Z, Zhao P, Li F, *et al*. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**:1047–57.
40. Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol* 2018;**63**:07TR01.
41. Maaten Lvd HG. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
42. Dao FY, Lv H, Wang F, *et al*. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 2019;**35**:2075–83.
43. Chen YY, Lin YH, Kung CC, *et al*. Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side Management for smart homes. *Sensors* 2019;**19**:2047.
44. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 1992;**46**:175–85.
45. Cui S, Zhao L, Wang Y, *et al*. Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. *Injury* 2018;**49**:1865–70.
46. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
47. Gao Y, Xi F, Zhang H, *et al*. Single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) in plants: the status of the bioinformatics tools to unravel the transcriptome complexity. *Curr Bioinform* 2019;**14**:566–73.
48. Lu CF, Hsu FT, Hsieh KL, *et al*. Machine learning-based radiomics for molecular subtyping of gliomas. *Clin Cancer Res* 2018;**24**:4429–36.
49. Cao R, Lopez-de-Ullibarri I. ROC curves for the statistical analysis of microarray data. *Methods Mol Biol* 2019;**1986**: 245–53.
50. Mazo C, Bernal J, Trujillo M, *et al*. Transfer learning for classification of cardiovascular tissues in histological images. *Comput Methods Programs Biomed* 2018;**165**:69–76.
51. Su QP, Zhao ZW, Meng L, *et al*. Superresolution imaging reveals spatiotemporal propagation of human replication foci mediated by CTCF-organized chromatin structures. *Proc Natl Acad Sci U S A* 2020;**117**:15036–46.
52. Marchal C, Sima J, Gilbert DM. Control of DNA replication timing in the 3D genome. *Nat Rev Mol Cell Biol* 2019;**20**:721–37.
53. Zhao PA, Rivera-Mulia JC, Gilbert DM. Replication domains: genome compartmentalization into functional replication units. *Adv Exp Med Biol* 2017;**1042**:229–57.
54. Gao F, Luo H, Zhang CT. DeOri: a database of eukaryotic DNA replication origins. *Bioinformatics* 2012;**28**:1551–2.
55. Rao SS, Huntley MH, Durand NC, *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.
56. Guillou E, Ibarra A, Coulon V, *et al*. Cohesin organizes chromatin loops at DNA replication factories. *Genes Dev* 2010;**24**:2812–22.