Exploiting epigenomic and sequence-based features for predicting enhancer-promoter interactions

Jianguo Zhou[1,a], Renyang Liu[2,b], Zifeng Wu[3,c], Jintao Zhang[4,d], Junhui Liu[5,e*]

[1]School of Software, Yunnan University, Kunming Yunnan China
Equal contributor
[2]School of Information Science and Engineering, Yunnan University, Kunming Yunnan China
Equal contributor
[3]School of Software, Yunnan University, Kunming Yunnan China
[4]School of Software, Yunnan University, Kunming Yunnan China
[5]School of Software, Yunnan University, Kunming Yunnan China

Abstract. How to discriminate distal regulatory elements to a gene target is challenging in understanding gene regulation and illustrating causes of complex diseases. Among known distal regulatory elements, enhancers interact with a target gene's promoter to regulate its expression. Although the emergence of many machine learning approaches has been able to predict enhancer-promoter interactions (EPIs), global and precise prediction of EPIs at the genomic level still requires further exploration.In this paper, we develop an integrated EPIs prediction method, called EpPredictor with improved performance. By using various features of histone modifications, transcription factor binding sites, and DNA sequences among the human genome, a robust supervised machine learning algorithm, named LightGBM, is introduced to predict enhancer-promoter interactions (EPIs). Among six different cell lines, our method effectively predicts the enhancer-promoter interactions (EPIs) and achieves better performance in F1-score and AUC compared to other methods, such as TargetFinder and PEP.

1 Introduction

Enhancers are key cis-elements that regulate spatiotemporal gene expression by contacting with their target genes. The existence of enhancers dramatically increases the complexity of regulatory networks in human and other organisms 1. Therefore, thousands of putative enhancers are mapped in mammalian genomes of different cell types, which outnumber coding genes 23. In many cases, one cognate gene can be controlled by multiple enhancers; in turn, one enhancer can also interact with more than one target gene 4. These create a complicated and nonlinear regulation network. Another aspect of increasing regulation network complexity lies in that enhancers are often located at a tremendous genomic distance away from cognate genes in mammalian and other vertebrate genomes.

In order to better understanding of EPIs, many investigations have been conducted to predict EPIs. Some algorithms use epigenomic features, such as TargetFinder 5, EpiTensor. By reconstructing regulatory landscapes from different features and integrating hundreds of genomics data sets, TargetFinder can accurately predict individual enhancer-promoter interactions using the features from enhancer, promoter, and window region between promoters and enhancer. TargetFinder revealed that the window region is more informative than the promoter and enhancer region; EpiTensor 13 proposes a novel unsupervised computational method to derive 3D interactions between distal genomic loci from 1D epigenomic data. While some only use a sequence as input data and can also gain a high performance to predict EPIs. SPEID is the first deep learning framework that only uses sequence features to predict the enhancer-promoter interaction 6. By integrating four features derived from the sequence, gene expression, and epigenomic features, IM-PET 8 predicts EPIs with a random forest classifier. PEP is an algorithm based on a boosted tree ensemble model to predict long-range EPIs. It consists of two modules (i.e., the PEP- motif and the PEP-word), which use different feature extraction methods. These researches also provide insight into how epigenetic features and sequences correlated to EPIs. Different from all the above methods, we develop a LightGBM-based algorithm to predict enhancer-promoter interactions named EpPredictor. Our result shows that the epigenomic factors-based features plus the sequence-based features, can reliably predict enhancer-promoter interactions and achieve better performance in F1-score and AUC when compared to TargetFinder and PEP.

* Corresponding author: [e] Corresponding author: hanks@ynu.edu.cn,[a]j.zhou@mail.ynu.edu.cn, [b]liurenyang@mail.ynu.edu.cn, [c]zifengwu@mail.ynu.edu.cn, [d]jintaozhang_0401@outlook.com

2 MATERIALS AND METHODS

2.1 Datasets

In this paper, we only use the enhancer-promoter pairs dataset from the TargetFinder 5, which include enhancer-promoter positive and negative sample pairs in six cell lines (GM12878, K562, IMR90, HeLa-S3, HUVEC, and NHEK). The distance between enhancer-promoter pairs is between 10 KB and 20 MB. The histone modification and transcription factor ChIP-seq datasets are from the ENCODE Project, including six cell lines. The histone and transcription factor datasets of each cell line contain peak and signal value features in their region.

By analyzing the TargetFinder method's datasets and the PEP method, we find that the positive and negative samples are quite different, and the negative samples are 20 times that of the positive samples. However, the unbalance of the positive and negative samples will lead to biased predictions, which result in overfitting. Therefore, we use all positive samples of the TargetFinder and randomly extract the same negative samples from the negative samples. The ratio between positive samples and negative samples is 1:1, thus solving the sample imbalance.

2.2 Determine the region of feature extraction

It is unclear which regions of the enhancer-promoter pairs within a chromosome are useful for predicting EPIs. Our dataset of enhancer-promoter pairs is from TargetFinder, which provides the enhancer's start point, the enhancer's endpoint, the promoter's start points, and the promoter's endpoint. Therefore, we define regions for extract features, as shown in Fig. 1:

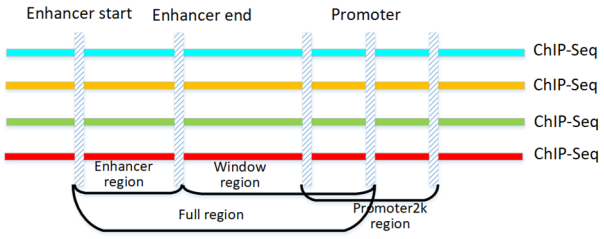

Fig1. Visualization of the location information of the chromosome.

Enhancer Region (ER): region from enhancer start point to enhancer end point.

Window Region (WR): region from the end point enhancer to start point of promoter.

Promoter 2k Region (PR): extend 2k from the right side and left side of the start point of promoter.

Full Region (FR): region from the start point of enhancer to the start point of promoter.

Note that the distance between the start point of promoter and the end point of promoter is mostly within 2k, and to obtain more useful information, we choose to extend 2k from the right side and left side of the start point of promoter. After defining the feature extracting regions, we map these regions location on chromosome and protein files, and get related sequences' and proteins' information.

2.3 Epigenomic features extraction

The features of the data are very essential to the algorithm model. A set of excellent features can well represent the information contained in the data, including invisible and explicit information. Thus, the machine learning model trained by excellent features can have better accuracy, generalization performance, and robustness. We extract multiple sets of features across six cell lines. By analyzing the distribution of enhancers and promoters on chromosomes in different regions at different signal values and peaks, we derive multi-group features based on protein signal value and peak. All four regions (i.e. ER, WR, FR, and PR) contains information about multiple sites of a protein. Therefore, it contains multiple sets of signal value and peak data.

2.4 Sequence feature extraction

Despite the epigenomic features, we also want to extract the sequence features from four regions (ER, WR, FR, and PR). Our sequence feature extraction method is the weight matrix, which is a motif descriptor. This method attempts to capture the inherent variability characteristics of the sequence pattern. It is usually composed of an equal number of sequences associated with a set of functional genes. Here we find that the length of each sequence we cut into L=20 is the best for predicting EPIs.

For example, here are the sequences of 3 eukaryotes. We tabulate the frequencies observed for each nucleotide at each position. The calculation of weight matrix as follows:

Table1. Calculation position weight matrix

Sequence1	AGTC
Sequence2	GTAC
Sequence3	AGCT

Table2. Position weight matrix

	1	**2**	**3**	**4**
A	2	0	1	0
T	0	1	1	1
G	1	2	0	0
C	0	0	1	2

Additionally, each number in this matrix represent the number of times that a given nucleotide has been observed at a given position. For example, nucleotide "A" has been observed in 2 aligned sequences at position 1 and is therefore represented as 2 in the matrix.

3 Result

3.1 Choose LightGBM as well as determine the region and feature selection

LightGBM (i.e., LGB) is higher efficiency, lower memory usage, and fast training speed algorithm. We do some extensive works and compare LightGBM with a support vector machine (SVC) 10 and Adaboost 11 under the epigenome and sequence features. Fig.2 shows that LightGBM achieves the highest F1-score compared to SVC and Adaboost in all six cell lines (i.e., K562, GM12878, IMR90, HeLa-S3, HUVEC, and NHEK). Different features and its related feature extraction regions should be carefully considered. We then evaluate LightGBM 12 with different features (i.e., DNA sequence feature, epigenomic factor feature, epigenomic+sequence feature) on four regions, which shows in Table 4 (i.e., ER, PR, WR, and FR).

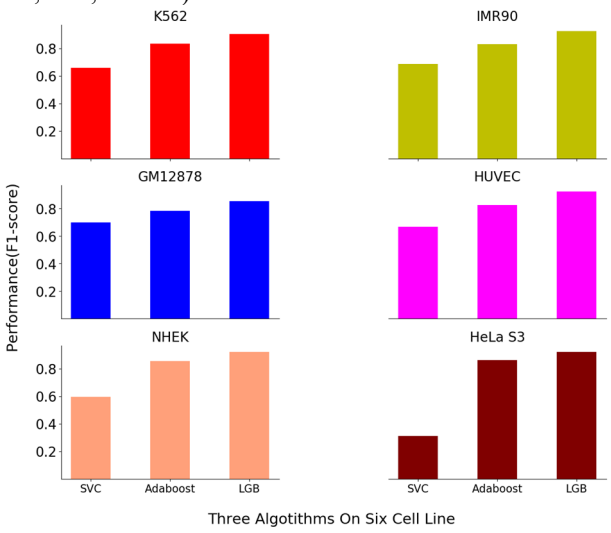

Fig2. Get the optimal algorithm map. In six cell lines, evaluation of F1-score using epigenomic plus sequence features in LightGBM, Adaboost, SVC algorithm

3.2 Compare EpPredictor to TargetFinder, and PEP

We use four regions features to evaluate the performance of EpPredictor across six cell lines compared to TargetFinder and PEP-integrate (Fig.3) in terms of F1-score AUC and MCC. The result shows that EpPredictor achieves better performance in the F1-score than the other two methods, named TargetFinder (on ER/PR/WR) and PEP (on ER/PR/WR). As details in Table 3, the average F1-score achieve by EpPredictor across six cell lines is 0.88, 5% higher than Target Finder, and PEP (on ER/PR/WR) (0.83 and 0.83). The F1-score is a significant improvement in all cell lines. In the K562 cell line, the F1-score of EpPredictor, TargetFinder (on ER/PR/WR), PEP (on ER/PR/WR) are 0.91, 0.85, and 0.82, respectively, which is increased 6% - 9% by EpPredictor. Also, in the GM12878 cell line, the F1-scores of EpPredictor is 0.85, which is 4% higher than TargetFinder.

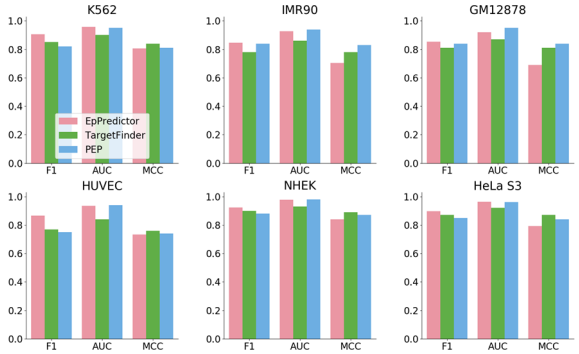

Fig3. Evaluation of EpPredictor on enhancer-promoter data from six cell lines (i.e. K562, GM12878, IMR90, HeLa-S3, HUVEC and NHEK) in comparison with TargetFinder and PEP methods in terms of F1-score AUC and MCC. Among them, green, blue, and pink represent the TargetFinder, PEP, and EpPredictor.

Table3. Detail performance evaluation of TargetFinder (on ER/PR/WR), PEP-Integrate (on ER/PR/WR) and EpPredictor in six cell lines.

Cell line	Method	F1-score	AUC	MCC
K562	TargetFinder	0.85	0.9	0.84
K562	PEP-integrate	0.82	0.95	0.81
K562	EpPredictor	0.91	0.96	0.81
GM12878	TargetFinder	0.81	0.87	0.81
GM12878	PEP-integrate	0.84	0.95	0.84
GM12878	EpPredictor	0.85	0.92	0.69
IMR90	TargetFinder	0.78	0.86	0.78
IMR90	PEP-integrate	0.84	0.94	0.83
IMR90	EpPredictor	0.85	0.93	0.70
HeLa-S3	TargetFinder	0.87	0.92	0.87
HeLa-S3	PEP-integrate	0.85	0.96	0.84
HeLa-S3	EpPredictor	0.90	0.96	0.79
HUVEC	TargetFinder	0.77	0.84	0.76
HUVEC	PEP-integrate	0.75	0.94	0.74
HUVEC	EpPredictor	0.87	0.94	0.73
NHEK	TargetFinder	0.90	0.93	0.89

NHEK	PEP-integrate	0.88	0.98	0.87
NHEK	EpPredictor	0.92	0.98	0.84

Table4. In six cell lines, LightGBM algorithm evaluates the size of F1-score in different combinations of epigenomic and DNA sequence features in the ER, WR, FR, and PR.

Region	features \ cell lines	K562	IMR90	GM12878	HUVEC	NHEK	Hela S3
ER	DNA sequences	0.5539	0.5189	0.5515	0.65	0.5353	0.7245
	epigenomic factors	0.547	0.5125	0.4624	0.5077	0.4825	0.4572
	epigenomic+sequence	0.5375	0.5112	0.5496	0.5928	0.5029	0.5344
PR	DNA sequences	0.699	0.7284	0.6838	0.594	0.8526	0.5611
	epigenomic factors	0.7419	0.634	0.7383	0.703	0.6501	0.7537
	epigenomic+sequences	0.7451	0.7264	0.7746	0.6913	0.8526	0.766
WR	DNA sequence	0.7438	0.7568	0.7494	0.6731	0.8336	0.75
	epigenomic factors	0.905	0.844	0.8587	0.8505	0.899	0.891
	epigenomic+sequences	0.9028	0.8418	0.8191	0.858	0.8336	0.8974
FR	DNA sequence	0.6359	0.639	0.6266	0.639	0.6981	0.6657
	epigenomic factors	0.9017	0.844	0.8587	0.8505	0.8972	0.891
	epigenomic+sequences	0.9044	0.8371	0.8616	0.8613	0.9172	0.8999
ER:Enhance Region, PR:Promoter Region, WR:Window Region, FR:Full Region.(In Sec 2.2)							

The enormous improvement observes in HUVEC, where the F1-scores of EpPredictor is 10%-12% higher than TargetFinder and PEP. As for the AUC, EpPredictor outperforms TargetFinder in the six-cell lines with about 4%-10% on AUC. and, EpPredictor similar PEP-interaction in three out of the six cell lines (HeLa S3, HUVEK, and NHEK) (0.96, 0.94, and 0.98, respectively)

The results in Table 4 indicate that LightGBM achieves better performance when selecting the epigenomic and sequence features. While the performance is not as expected if only selecting DNA sequence features. From Table 4, we observe that in five cell lines (i.e., GM12878, K562, HeLa-S3, HUVE. For instance, in GM12878, LightGBM reaches 0.8616, which is the best in this column. The only exception is in IMR90, where it got 0.8371, which is slightly lower than 0.8418 under the window region. The full region (FR) is more informative than the other three regions when incorporating epigenomic and sequence features (Table 2).

Interestingly, although EpPredictor achieves the higher F1-score and AUC, the performance of MCC is not as expected. In particular, in IMR90, TargetFinder (on ER/PR/WR) and PEP (on ER/PR/WR) achieve 0.81 and 0.84 MCC, achieve better performance as compared to EpPredictor (0.69). One of the classifiers may have a higher F1-score value and a lower MCC value, which means that a single score cannot measure all the classifier's advantages and disadvantages. In summary, the EpPredictor has a clear advantage over the TargetFinder and PEP in F1-score. This trend also maintains in the AUC score.

4 Discussion and Conclusion

We developed EpPredictor to predict EP interaction based on LightGBM. Compared to other models such as SVM and Adaboost, LightGBM achieves higher performance. Many methods only use the features from the promoter region and enhancer region to predict EPIs. While TargetFinder suggests that the window region's features are more critical to predict EPIs [13,14,15]. Our methods revealed that the full region with sequence and epigenetic features is more efficient in predicting EPIs than the window region. In order to get better performance, EpPredictor integrates epigenomic features which are extracted from the four regions (ER/PR/PR/FR) according to the information of histone and transcription factor, and sequence features which are also extracted from the four regions (ER/PR/PR/FR).

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61762089, Grant 91631305, Grant 61863036, Grant 61862067 and Grant 81560380 and in part by the Science and Technology Innovation Team Project of Yunnan Province under Grant 2017HC012.

REFERENCE

1. Schoenfelder, S., Fraser, P.: Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics, 1 (2019)
2. Shen, S., Madau, P., Aguirre, A., Guedes, J., Mayer, L., Wadsley, J.:The origin of metals in the circumgalactic medium of massive galaxies at z= 3. The Astrophysical Journal 760(1), 50 (2012)
3. Ecker, J.R., Bickmore, W.A., Barroso, I., Pritchard, J.K., Gilad, Y., Segal, E.: Genomics: Encode explained. Nature 489(7414), 52 (2012)
4. Osterwalder, M., Barozzi, I., Tissi` eres, V., Fukuda-Yuzawa, Y.,Mannion, B.J., Afzal, S.Y., Lee, E.A.,

Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al.: Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554(7691), 239 (2018)

5. Whalen, S., Truty, R.M., Pollard, K.S.: Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature genetics 48(5), 488 (2016)
6. Singh, S., Yang, Y., Poczos, B., Ma, J.: Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. bioRxiv, 085241 (2016)
7. He, B., Chen, C., Teng, L., Tan, K.: Global view of enhancer–promoter interactome in human cells. Proceedings of the National Academy of Sciences 111(21), 2191–2199 (2014)
8. Yang, Y., Zhang, R., Singh, S., Ma, J.: Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics 33(14), 252–260 (2017)
9. Yang, Y., Zhang, R., Singh, S., Ma, J.: Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics 33(14), 252–260 (2017)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
11. Freund, Y., Schapire, R., Abe, N.: A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence 14(771-780), 1612 (1999)
12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp. 3146–3154 (2017)
13. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J.W., Ding, B., Li, N., Zheng, L., Wang, W.: Constructing 3d interaction maps from 1d epigenomes. Nature communications 7, 10812 (2016)
14. Zeng, W., Wu, M., Jiang, R.: Prediction of enhancer-promoter interactions via natural language processing. BMC genomics 19(2), 84 (2018)
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp. 3146–3154 (2017)