# COMPUTATIONAL METHODS TO IDENTIFY THE TARGET GENES OF THE ENHANCERS

A PREPRINT

**Jiaxin Wang**
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213
jiaxiwa@andrew.cmu.edu

**Xi Xu**
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213
xix@andrew.cmu.edu

**Shili Wang**
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA 15213
shiliw@andrew.cmu.edu

May 13, 2021

## 1 Introduction

Enhancers are cis-regulatory elements on DNA sequences that affect the level of gene expression, usually locate far from the transcription start site and have high degrees of tissue specificity. [1] They regulate gene expression by interacting with their target gene promoters on the chromosomes by forming the ring structures and recruiting multiple transcription factors binding to the regions to initialize transcription. Due to the distances to gene sites and variable control characteristics, predicting the enhancer- gene interactions is a big challenge and of great importance.

In recent years, with the increasing genome-wide experimental data and development of bioinformatics, a series of computational prediction methods based on different enhancer features have been developed. Most popular methods are combined with various genomic information including chromatin accessibility, tRNA, histone marks, etc, [2,3,4,5,6,7,8] that can be used in gene regulation in different tissues and pave the way of human diseases such as cancer, Alzheimer, etc. [5,8,9,10,11].

One widely used approach for predicting the relationships between enhancers and gene regulation is finding the correlation of epigenomic information between enhancers and genes across multiple biosamples. Ernst J et al [2] firstly examined histone modification data near enhancer in 9 human cell lines. By associating analysis of RNA-seq expression data of genes within a distance of 125 kb, they searched for enhancer-gene pairs with common variation patterns to predict the action sites of enhancers. Then others followed the idea to find the enhancer-gene interactions by relating multiple genomic signals. Andersson et al. [12] used Cage data to predict the action sites of enhancer. Prestige and PreSTIGEouse were developed by Corradin et al. [13] and Factor et al. [14] to predict enhancer action sites in humans and mice, respectively. He et al. [15] developed IM-PET, which uses a combination of multiple genetic traits to predict the action sites of enhancer. Moreover, ABC method [21] characterizes the interaction frequency of enhancer and promoter pair, and predicts the pairs combining the active signals of enhancers.

Other methods such as supervised learning models aim to predict the interactional pairs based on the already discovered enhancer-gene datasets. Among them, PEP method [16] integrates PEP-motif and PEP-word models to predict enhancer-promoter interactions by extracting sequence features from enhancer and promoter sites of specific cell types.The PetModule method [19] is also based on motifs, that uses similar properties like IM-PET. TargetFinder aligns the sequence based on boosted trees. With the rapid development in deep learning techniques[17], SPIEID method [18] applied deep neural networks (DNN) to predict enhancers according to the sequence characteristics of enhancers and promoters. Epiann [20] uses attention-based neural network model that pay more attention to the inner characteristics and make more accurate predictions.

Although these approaches have successfully identified the subcollections of potential pairs, the methods still yet lack of systematical evaluation and comparison. In this project, we aim to use benchmark datasets to test several published

computational methods for linking enhancers with genes, including distance method, PEP and ABC methods. Then we can compare different methods and evaluate how distances of pairs changes their performances.

## 2 Methods

Code required to reproduce the results are available at `https://github.com/wwwsl0727/02710-group-project`.

### 2.1 Supervised Method(PEP-motif method)

Here we test one supervised method named PEP-motif method with the HiC dataset. The workflow for the PEP-motif method is described in Figure 1. Firstly we use FIMO software to finish motif search step, then we get a form as ($l_1^{(i)}$, $l_2^{(i)}$,... $l_M^{(i)}$). The jth element $l_j^{(i)}$ represents the number of occurrence for the j-th motif for the i-th sequence. Then after the motif normalization step, we generate the feature vector of the i-th sequence as $f^{(i)} = (f_1^{(i)}, f_2^{(i)}, ..., f_M^{(i)})$, where $f_m^{(i)} = l_m^{(i)}/L_i$, m = 1, ..., M, $L_i$ is the length of the i-th sequence. To formulate the motif features for the enhancer-tss pair, we concatenate the motif feature vectors for both the enhancer and tss. So the input vector is $(enhancerf_1^{(i)}, enhancerf_2^{(i)}, ..., enhancerf_M^{(i)}, tssf_1^{(i)}, tssf_2^{(i)}, ..., tssf_M^{(i)})$, i = 1, ..., the length of train pair, in which $enhancerf_j^{(i)}$ means the value for the i-th motif in the enhancer of the j-th enhancer/tss pair and $tssf_1^{(i)}$ means the value for the i-th motif in the tss of the j-th enhancer/tss pair.

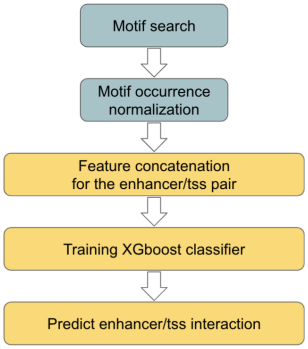


Figure 1: the workflow for the PEP-motif method

XGBoost belongs to ensemble methods, that builds a strong classifier via creating multiple classifiers and combining the predictions of all these classifiers. So the final model tends to be more generalizable and robust compared to a single classifier and it allows the model to learn subtle and deep interactions between features. As the model is designed to optimize its computing workload as well as its memory allocation, XGBoost is fast. Here We introduce the XGBoost learning library to implement XGBClassifier to make prediction if the TSS is the target region of the enhancers based on feature vectors generated by PEP-Motif.
To use our model to predict whether a given tss is the target gene of the enhancer. The input is a motif feature vector of the tss/enhancer pair. The output is a label which indicates whether the tss and enhancer have association or not.

### 2.2 ABC model

There are 3 steps to get the ABC score. Firstly, the chromatin accessibility is calculated by calling peaks on the first replicate of DNase-seq data. Each candidate element is classified as a promoter, genic or intergenic element. The promoters are regions that are within 500bp of any annotated TSS. The distal elements include genic or intergenic element. Secondly, the enhancer of activity is calculated from DNase-seq and H3K27ac ChIP-seq reads in the candidate enhancer regions. A denotes the genomic mean of read counts of DHS and H3K27ac chromatin immunociprecipitation at element E.[19] C denotes the contact frequency between the E and the promoter of gene G with 5kb resolution. Finally, ABC scores are calculated by combining activities and Hi-C interaction frequencies.

If the ABC score between a gene and a enhancer is large, it's more likely that the gene is the target of the enhancer.

$$ABC_{score_{E,G}} = \frac{A_E * C_{E,G}}{\sum_{all\ elements\ with\ 5\ mb\ of\ the\ tss} A_e * C_{e,G}}$$

### 2.3 Unsupervised method

The distance method is used to predict the score based on the inverse of the distance between the cCRE-TSS pairs. There is no parameter to tune in this model.

## 3 Implementation Details

### 3.1 Datasets

The datasets we use in the unsupervised learning model are RNAII CHIA-PET dataset and HI-C dataset of GM12878 cell lines which contain 3d chromatin interactions.

The link given in the reference paper[19] doesn't contain the original dnase data of GM12878 for the ABC model. Also, due to time limitation, it's hard to find the original .bam files with peak information online.

Another difficult thing is it's hard to find the true label of the crisper-disturbed dataset.

As a result, we used the k562 crisper disturbed benchmark dataset and used the allpreditions.txt from the paper[19] with ABC score as the input.

### 3.2 Benchmark Generation

The pipeline of generating benchmark datasets and making evaluation on three methods is illustrated in Figure 2. The datasets we use are 3D chromatin interactions, which include ChIA-PET and Hi-C interactions.

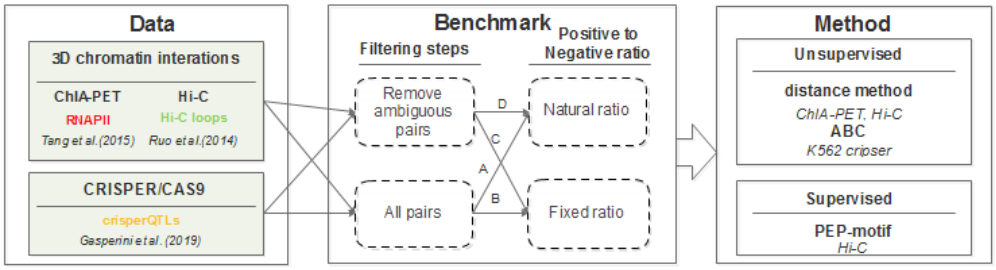


Figure 2: Pipeline

The links with one locus site in the range of distal enhancer and the other falling within 2kb of annotated TSS are selected as the input. Since 75% of the 3D chromatin interaction sites overlap more than one gene, ambiguous pairs are removed in order to improve prediction accuracy.

Besides, the positive enhancer-gene pairs mean the enhancers promote the transcriptions and negative pairs mean they suppress the transcriptions. Since the natural ratio of negative pairs to positive pairs is very high, negatives pairs are randomly discarded. As a result, one dataset with one-fourth positive-negative ratio is generated.

Therefore, we divided the integrated dataset to form four subsets, by whether filtering the ambiguous pairs or keeping the natural ratio.True labels are pairs from the original linked dataset with enhancers and tss. Then the benchmark data are used as the input for distance, PEP, and ABC methods.

### 3.3 Feature extraction in PEP-Motif

The strategy for motif feature extraction is to use a software FIMO to scan reference motif information for sequence. A P-value threshold of 1e-04 is applied to identify the result for FIMO match, then compute the normalized motif occurrence for enhancer/TSS. Assume $l_1^{(i)}$, $l_2^{(i)}$, $l_3^{(i)}$ are the number of occurrence for the specific motif for the i-th sequence, then the feature vector of the i-th sequence is $f^{(i)} = (f_1^{(i)}, f_2^{(i)}, ..., f_M^{(i)})$, where $f_m^{(i)} = l_m^{(i)}/L_i$, m = 1, ..., M, $L_i$ is the length of the i-th sequence. The reference motif we used is the motifs for HUMAN core transcription factors

which are primary binding patterns and been verified to be robustly indicating binding site across multiple experiments. The size of the reference motif is 401 and we download the reference motif from Homo sapiens Comprehensive Model Collection (HOCOMOCO) database. To formulate the motif feature for the enhancer-tss pair, a concatenated form vector based on the feature vectors of the enhancer region and tss region is generated.[18]

### 3.4 Parameter tuning in XGBoost

We perform parameter tuning in XGBoost in order to find a balance in a high AUC as well as avoid overfitting. We set parameter max_depth, learning_rate, n_estimators as well as nthread manually and set all other parameter as default.
For the max_depth parameter which control the depth of the tree, a greater depth always enables the model to capture more complex interactions between features but it also faces a challenge of overfitting. Finally, we set the max depth equals to 10.
Learing_rate parameter controls the step size to take in each boosting step.We set this parameter to 0.1 in order to allow XGBoost to take more conservative step and it help to prevent overfitting.
We set the n_estimators parameter from default 100 to 1000, which enables the number of trees in the ensemble to be 1000. Such setting can improve the performance of the model while spending more time on training.
nthread as 50 allows the parallel way to run the XGBoost model.

## 4 Results and Conclusions

### 4.1 Results of unsupervised method

#### 4.1.1 AUC comparisons between 2 datasets for the distance based method

Here are the roc curves for applying distance method on 2 different datasets. The roc area for Hi-c is 0.79 and the roc area for CHIA-PET is 0.85. It means that the distance method performs better in CHIA-PET dataset.

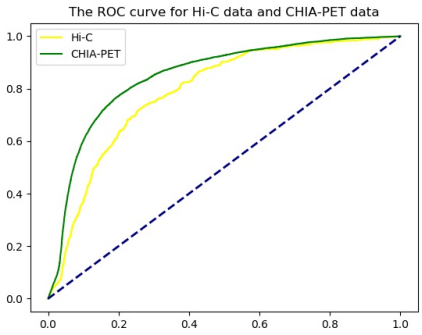


Figure 3: the roc curve for distance method based on 2 datasets

#### 4.1.2 The relationship between the distance distribution and the roc value for distance method

In order to get the performance for how distance method performs in pairs with relative long distance, the distance threshold is used to filter the pairs whose distance is less than the threshold.

Firstly, the distribution of the distance of all pairs for 2 datasets are in figure 4 and figure 5. The distribution is nearly uniform, In other words, if using 2,000,000 as the threshold for Hi-c dataset, 20% of the whole dataset will be removed.

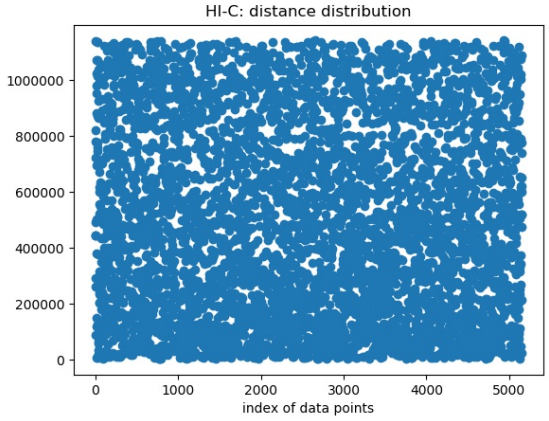


Figure 4: the distance distribution for Hi-C dataset

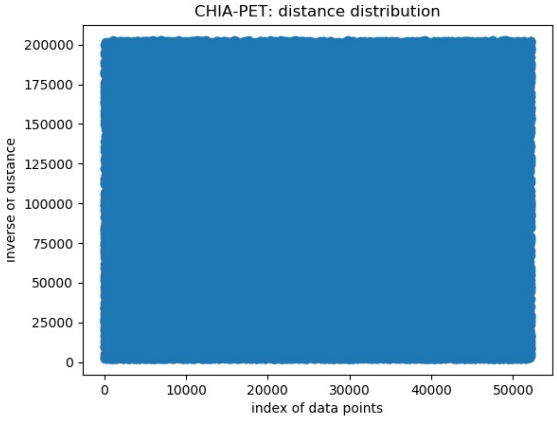


Figure 5: the distance distribution for CHIA-PET dataset

Figure 6 and figure 7 show that when the thresholds increase, the auc decreases nearly linearly. It means that the distance method performs worse in pairs with long distance.

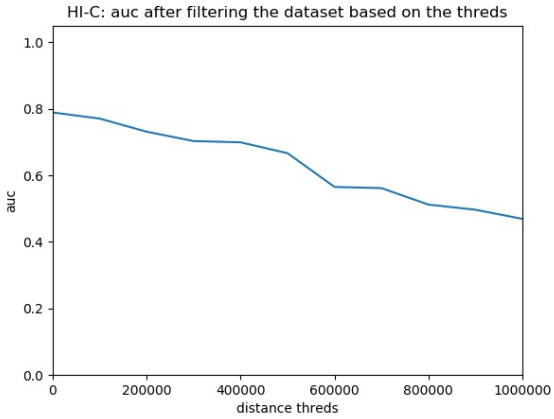


Figure 6: the relationship of roc and the distance threshold for Hi-C dataset

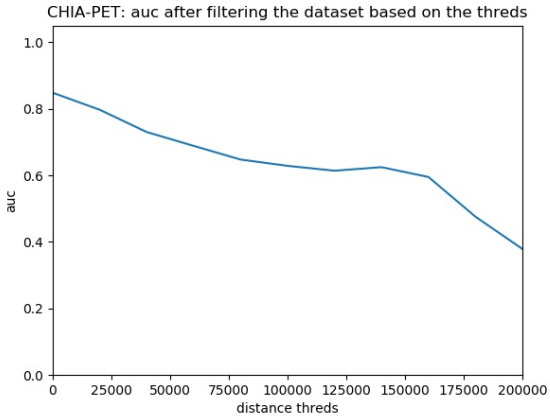


Figure 7: the relationship of roc and the distance threshold for CHIA-PET dataset

### 4.2 Comparison between PEP-motif method and unsupervised method

The performance of the models all measured by the Area Under the ROC Curve(AUC). An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). A higher AUC always indicate a better performance of the model. For HiC dataset the AUC for PEP-motif method is 0.76 which has slight difference between the AUC(0.79) for distance method for the same dataset(Figure 8, Figure 9).

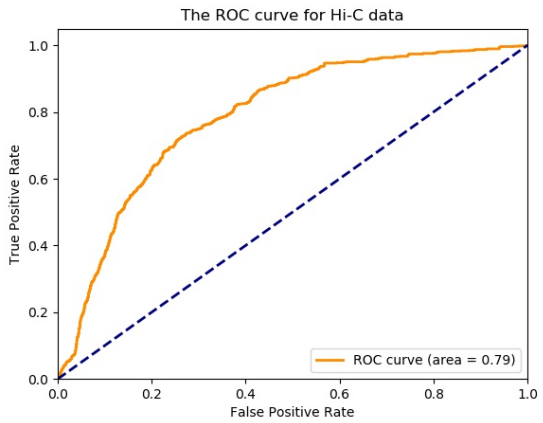


Figure 8: the roc curve for distance method based on the HiC dataset

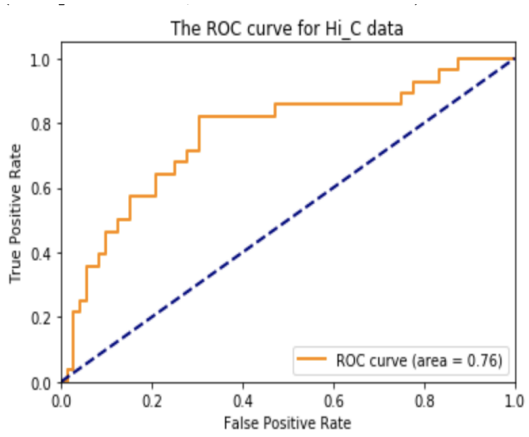


Figure 9: the roc curve for PEP-motif method based on the HiC datase

To have a better measurement for these two methods. We test the performance for these two methods in a biased dataset. The biased dataset contains approximately 500 cCRE-TSS pairs, we extracted these pairs from the original 5155 pairs HiC-dataset based on the standard to choose the first 10% pairs ranked by distance. All these pairs has distance > 100,000, and the pairs with label 1 only account for 2% for these 500 pairs.
The AUC for the distance method drops down to 0.47 as it shows on figure 10 but AUC for PEP-motif method is 1 in figure 11 here which indicates it can predict every pair in the test dataset correctly.

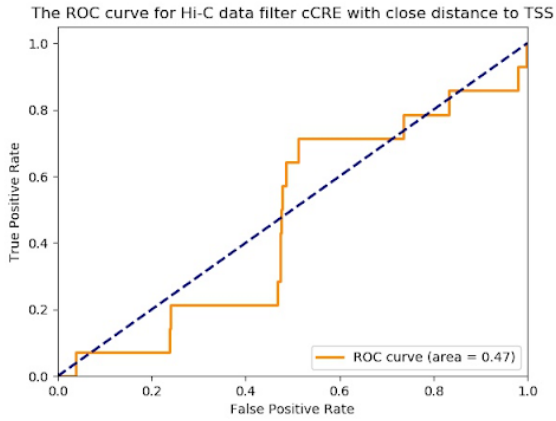


Figure 10: the roc curve for distance method based on the biased dataset

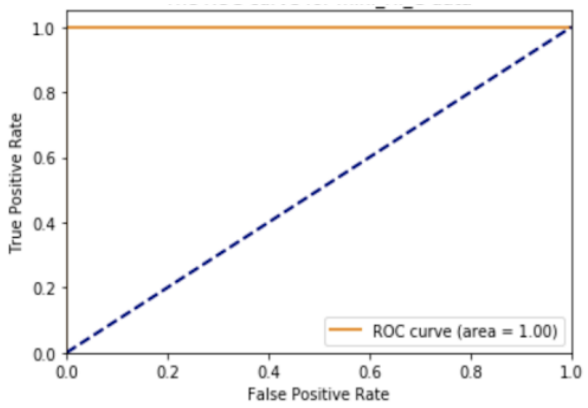


Figure 11: the roc curve for PEP-motif method based on the biased dataset

Such phenomenon is very likely to be caused by the current pairs with long distances. The distance method extracts only distance-related features. However, PEP-motif method extracts transcription factor binding site (TFBS) motifs information. For the long distance-pairs, the distance features may be misleading since the larger the distance, the lower association probability is in the distance method. However, in the real life, it is obvious not the current case.

Moreover, the pairs with long distances don't impact the progress to extract motif features. It is very likely that specific transcription factor binding site (TFBS) motifs plays more important rule in predicting the long-distance pairs than the distance features. So PEP-motif method has a good performance for the biased dataset while distance method doesn't.

### 4.3 Comparison between ABC model and the unsupervised method

For the full k562 crisper disturbed dataset, the auc for ABC model is 0.78 and the auc for the distance method is 0.86 as in figure 11 and figure 12.

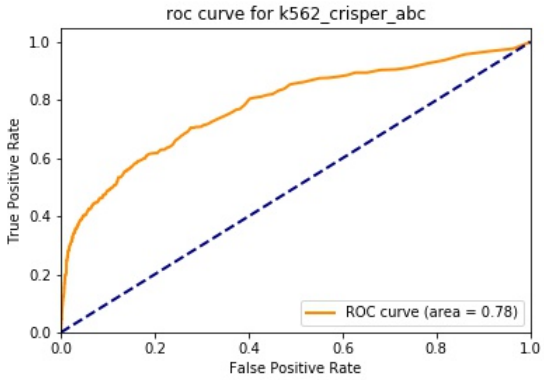


Figure 12: roc curve for ABC model based on the k562 crisper disturbed datasets

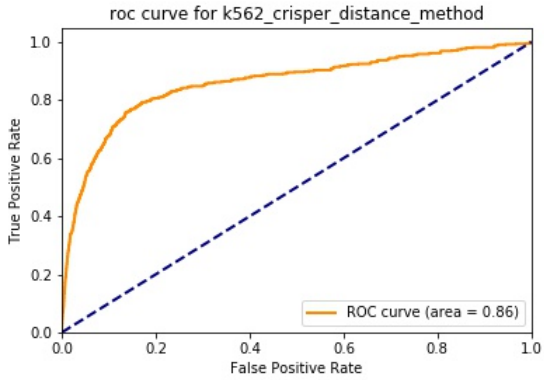


Figure 13: roc curve for distance method based on k562 crisper disturbed datasets

However, using 500,000 as the distance thredshold to filter $71.4\%$ of all pairs. The auc for ABC model is 0.53 and auc for distance method is 0.51. The ABC works slightly better than the distance method in this case. However, compared with the full dataset, the auc for ABC model drops a lot. The reason is that the default distance that a enhancer interacts with its target tss is less than 500,000. When filtering all the pairs with distance less than 500,000, ABC model works terribly.

It's important for ABC model to choose a valid distance threshold.

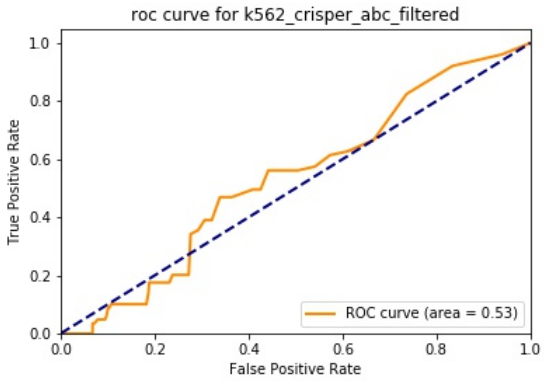


Figure 14: roc curve for ABC model based on the filtered k562 crisper disturbed datasets

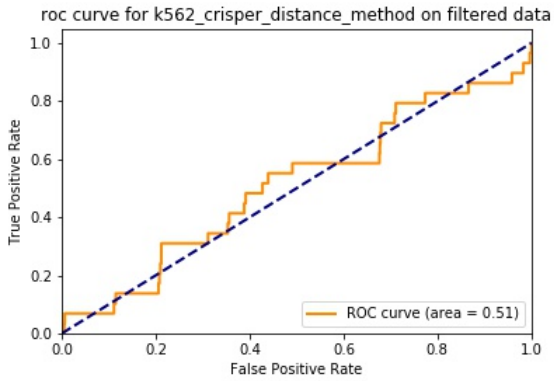


Figure 15: roc curve for distance method based on filtered k562 crisper disturbed datasets

### 4.4 Limitation and further direction

The goal to measure the performance of different models is to choose which model to use in a specific dataset. For the current two methods, we observed that distance method is very straightforward to implement while the performance will drop down when dealing with long-distance pairs. So the next step is to find a distance threshold which helps to make decision whether we should implement PEP-motif method or use distance method.
The biased dataset we generated is a relative small dataset and may lead to the measurement of different method not very persuasive. We will measure the performance for our supervised and unsupervised method on a relative big dataset which consists of long-distance pairs.
The threshold of ABC should be carefully selected in order to get a better performance than the distance method. Finally, there exists other models which are very useful to detect the enhancer-target gene pairs, we should take a close look at these models and compare with the models we have explored too.

## References

[1] CORRADIN O SCACHERI P C. Enhancer variants Evaluating functions in common disease[J]. Genome Medicine 2014 6(10) 85.

[2] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

[3] ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74

[4] Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6.

[5] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

[6] He Y, Gorkin DU, Dickel DE, Nery JR, Castanon RG, Lee AY, et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proc Natl Acad Sci U S A. 2017;114:E1633–40.

[7] Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.

[8] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.

[9] Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59

[10] Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014;518:337–43.

[11] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genet. 2015;47:1228–35.

[12] ANDERSSON R GEBHARD C MIGUEL-ESCALADA I et al. An atlas of active enhancers across human cell types and tissues[J]. Nature 2014 507(7493): 455.

[13]CORRADIN O SAIAKHOVA A AKHTAR-ZAIDI B et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits[J]. Genome Research 2014 241 5.

[14] FACTOR D C CORRADIN O ZENTNER G E et al. Epigenomic comparison reveals activation of "seed" enhancers during transition from naive to primed pluripotency[J]. Cell Stem Cell 2014 146 854.

[15] HE Bing, CHEN Changya, TENG Li, et al. Global view of enhancer-promoter interactome in human cells[J]. Proceedings of the National Academy of Sciences of the United States of America 2014 111(21): E2191.

[16] Yang Yang, Ruochi Zhang, Shashank Singh, Jian Ma, Exploiting sequence-based features for predicting enhancer–promoter interactions, Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages i252–i260

[17] Lappalainen T, Sammeth M, Friedländer MR, PAC 'tH, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.

[18] Fulco, C.P., Nasser, J., Jones, T.R. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet 51, 1664–1669 (2019). https://doi.org/10.1038/s41588-019-0538-0

[19] Zhao C, Li X, Hu H: PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep 2016, 6:30043.

[20] Singh S, Yang Y, Póczos B, Ma J: Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quantitative Biology 2019, 7(2):122-137.

[21] Mao W, Kostka D, Chikina M: Modeling Enhancer-Promoter Interactions with Attention-Based Neural Networks. bioRxiv 2017:219667.

[22] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA et al: Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet 2019, 51(12):1664-1669.

## 5 Supplementary

The barcharts of enhancer-gene predicitions in four benchmark datasets are illustrated as Figure 16. CHIA-PAT and HI-C are included. Figure 16(a) is positive va all and Figure 16(b) is the ratio.

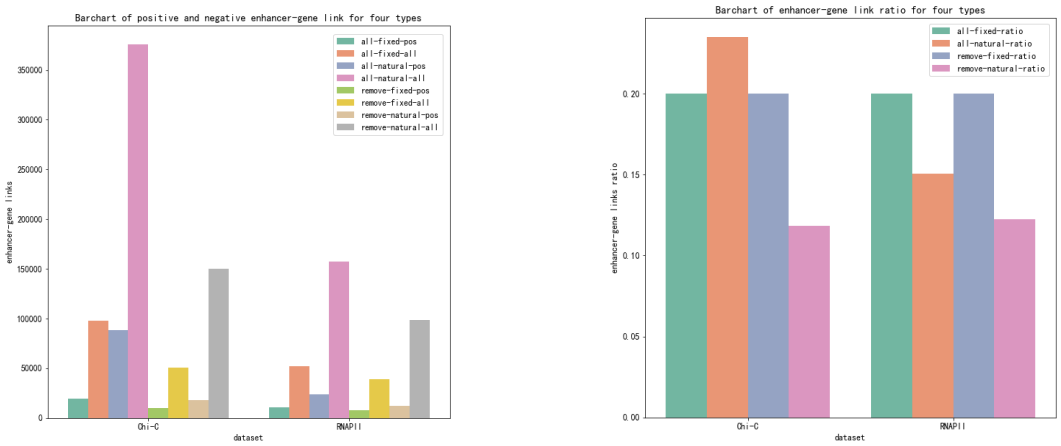


Figure 16: Benchmark Statistic