論 文 の 内 容 の 要 旨

論文題目　A new computational method to predict transcriptional activity of a DNA sequence from diverse datasets of massively parallel reporter assays
（多様な超並列レポーターアッセイデータセットを用いたDNA配列の内在的転写活性予測に関する新規計算手法の開発）

氏　名　劉 瑩

Introduction

Gene transcription regulatory code is surely encoded in the sequences of cis-regulatory elements and revealing functional elements from *cis*-regulatory elements is a key of exploring its regulatory mechanisms of transcription. Massively parallel reporter assay (MPRA) technology is kind of reporter assay based on DNA barcoding and next generation sequencing. The application of MPRAs for different purposes produced a large body of data which contain the sequence primary activities. To detect the functional sequences, it requires a computational model to estimate the relationship between sequences and transcription activities. However, a computational method which could be applied to diverse MPRA data sets is not existed yet. In this research, I designed a computational method to predict transcription activities using sequences and the corresponding activities by TRANSFAC database and machine learning algorithms of regression tree and MARS. According to analysis of predictive functions which estimated by the proposed method, it could reveal the active transcription factor binding sites (TFBSs). The proposed method could be applied to diverse MPRA as well as to luciferase reporter assay data sets despite different transfected cell types, different sequence lengths (several ten bp to more than 1k bp), different number of sequences (several hundred to more than several ten thousand) and different sequence types (promoters, enhancers, artificial sequences, ChIP-seq peaks and genomic variants). The applications of the proposed method also suggest that the method could predict the transcription activities of unknown sequences by using the predictive functions for known MPRA data sets.

Material and Methods

The proposed method consists of four steps: 1. Data pre-processing to format different MPRA data sets; 2. TRANSFAC database searching to encode sequence into variables and construct the explanatory variable matrix; 3. Variables clustering to assemble variables into more compact subpopulation by regression tree; 4. Perform MARS in different clusters to construct predictive

functions (Figure 1).

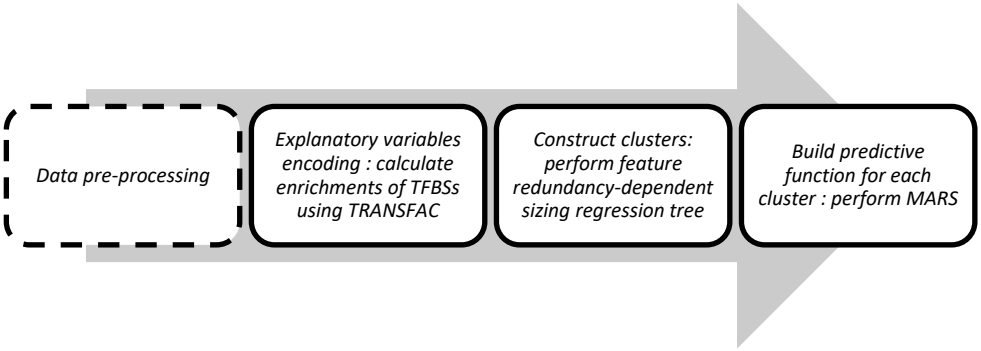

Figure 1. Workflow of proposed model

Data sets

To demonstrate the usability, we applied the proposed method to 10 public data sets from 8 previous works (Table 1) that contain 8 MPRA data sets and 2 luciferase reporter assay data sets.

Data sets	Description	Construct lengths	Cell types	Assayed loci	#Constructs
CREInducedInHEK293	CRE enhancer with 10% random mutations	87bp	HEK293	*ex vivo*	27000
DHSInMouseRetina	3500 DNase I hypersensitive sites	181-703bp (median 466bp)	mouse retina	*ex vivo*	27161
TFBS75InYeast	Designed 75 yeast TFBSs	103bp	yeast	*in vivo*	6016
TFBS12InHepG2Mouse	12 liver-specific TFBSs assayed in HepG2 and Mouse	168bp	mouse,HepG2	*in vivo, ex vivo*	4742
RBCVariantsGATA1InK562	2,756 SNPs assayed in GATA1 overexpression+/- K562	145bp	K562	*ex vivo*	15733
PromoterLucInHEK293	Promoters	755-1201bp (median 1081bp)	HEK293	*ex vivo*	734
CREBBPInMouseNeuron	253 distal enhancers and 234 promoters assayed by MPRA and STARR-seq	139bp	mouse cortical neurons	*ex vivo*	3409
PromoterLuc8celltypes	Promoters assayed in 8 cell types	614-1301bp (median 983bp)	Ags G402 HCT116 Hela Hepg2 HT1080 T98G U87mg	*ex vivo*	4575

Table 1. The basic information of data sets.

Performances of the proposed method

The proposed method was applied to the 10 data sets (Table 1) and obtained the predictive precisions (Pearson's R of predicted values and experimental values) were approximately 0.5 to 0.9 (table 2). The open tests were estimated by 100-fold cross-validation and obtained the similar predictive precisions with close test. The number of predictors were also small which are generally lower than 50.

data set	*close test*	*open test*	*# of predictors*
RBCVariantsGATA1InK562	0.55	0.49	16 - 30
RBCVariantsCtrlInK562	0.57	0.50	21 - 26
CREBBPInMouseNeuron	0.64	0.50	21 - 36
DHSInMouseRetina	0.64	0.52	20 - 48
TFBS12InHepG2	0.73	0.71	28 - 28
PromoterLuc8celltypes	0.73	0.70	21 - 50
TFBS12InMouse	0.78	0.76	35 - 35
CREInducedInHEK293	0.83	0.81	25 - 47
PromoterLucInHEK293	0.92	0.85	28 - 28
TFBS75InYeast	0.92	0.91	16 - 30

Table 2. The number of predictors of the estimated predictive functions and the correlation coefficients between predicted values and experimental values of close test and open test.

Application

From the analysis of “CREInducedInHEK293” and “CREBBPInMouseNeuron” , the candidate-active TFBSs estimated by the proposed method are characterized via tree structure. The TFBSs tree have the advantages of simply understandable and could provide the biological information for data set clustering (Figure 2).

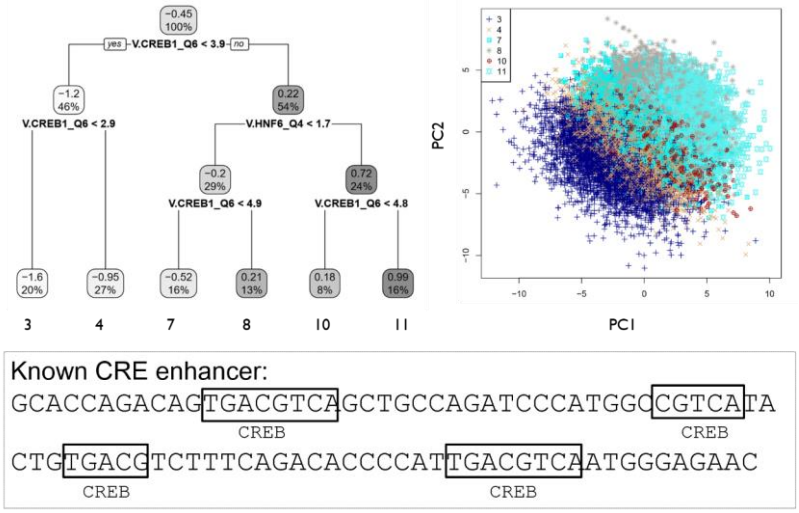

Figure 2. (Top left) Candidate-active TFBS trees for “*CREInducedInHEK293*” data sets. The values shown in each cluster indicate the average activity among samples within the corresponding cluster, and the percentages represent the sample proportions in the cluster. (Top right) PCA plot of “*CREInducedInHEK293*” with different colors indicate the different cluster showed in the candidate-active TFBS tree. (Lower)Known TFBSs in CRE enhancer.

In the analysis for the data sets “*RBCVariantsCtrlInK562*” and “*RBCVariantsGATA1InK562*”, the proposed method could pick up the candidate-active TFBSs that response to GATA1 OE. The candidate-active TFBS tree and the sample clusters estimated by the proposed method also show different structure of different experimental conditions. It suggests that the proposed method could detect the experimental condition-specific TFBSs (Figure 3).

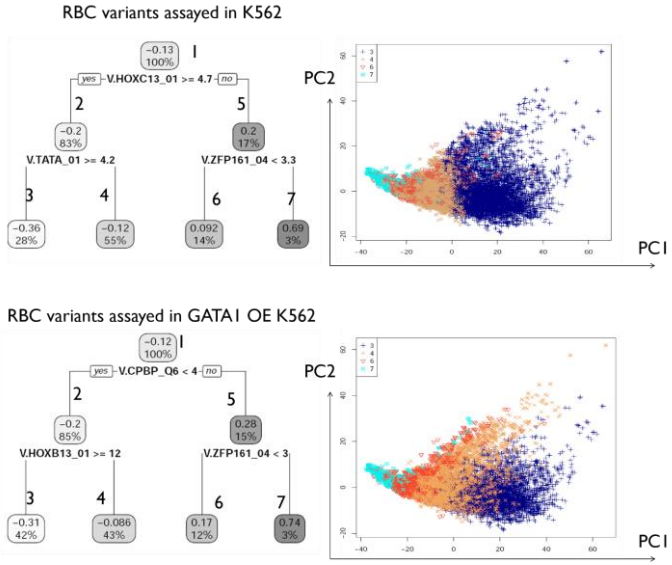

Figure 3. (Left) Candidate-active TFBS trees for data sets "*RBCVariantsCtrlInK562*" and "*RBCVariantsGATA1InK562*" data sets. The values shown in each cluster indicate the average activity among samples within the corresponding cluster, and the percentages represent the sample proportions in the cluster. (Right) PCA plot of sets "*RBCVariantsCtrlInK562*" and "*RBCVariantsGATA1InK562*" with different colours indicate the different cluster showed in the candidate-active TFBS tree.

Using the predictive functions for data set of "*CREBBPInMouseNeuron*," I predicted new sequences of 18 motifs with the same experimental condition as "*CREBBPInMouseNeuron*." For the data set of "*PromoterLucInHEK293*", I predicted the transcription activities by the predictive functions estimated for "*PromoterLuc8celltypes*" (Figure 4). This suggests that the transcription activities of unknown sequences could be predicted by known data set in some extent, despite of the cell types.

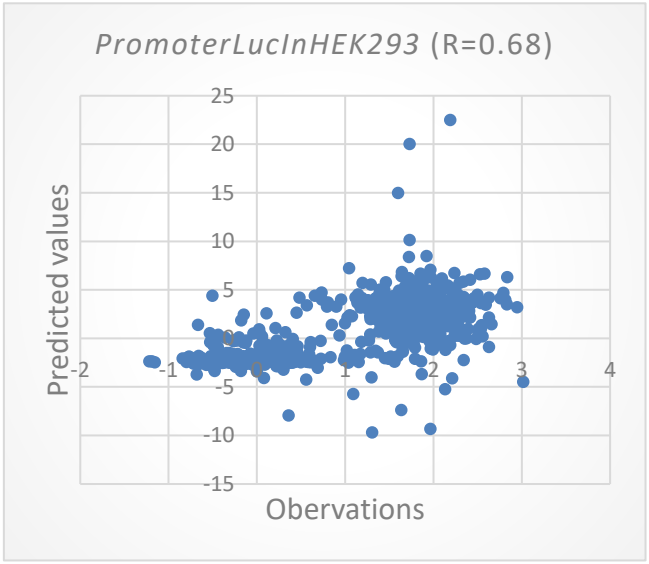

Figure 4. Plots between predicted transcription activities of "*PromoterLucInHEK293*" that were estimated by predictive functions of "*PromoterLuc8celltypes*" and observations of "*PromoterLucInHEK293*."