He *et al. BMC Systems Biology* 2018, **12**(Suppl 4):44
https://doi.org/10.1186/s12918-018-0570-1

BMC Systems Biology

RESEARCH Open Access

70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features

Wenying He[1], Cangzhi Jia[2], Yucong Duan[3] and Quan Zou[1*]

From The 11th International Conference on Systems Biology (ISB 2017)
Shenzhen, China. 18-21 August 2017

Abstract

Background: Promoter is an important sequence regulation element, which is in charge of gene transcription initiation. In prokaryotes, σ^{70} promoters regulate the transcription of most genes. The promoter recognition has been a crucial part of gene structure recognition. It's also the core issue of constructing gene transcriptional regulation network. With the successfully completion of genome sequencing from an increasing number of microbe species, the accurate identification of σ^{70} promoter regions in DNA sequence is not easy.

Results: In order to improve the prediction accuracy of sigma70 promoters in prokaryote, a promoter recognition model 70ProPred was established. In this work, two sequence-based features, including position-specific trinucleotide propensity based on single-stranded characteristic (PSTNPss) and electron-ion potential values for trinucleotides (PseEIIP), were assessed to build the best prediction model. It was found that 79 features of PSTNP_{SS} combined with 64 features of PseEIIP obtained the best performance for sigma70 promoter identification, with a promising accuracy and the Matthews correlation coefficient (MCC) at 95.56% and 0.90, respectively.

Conclusion: The jackknife tests showed that 70ProPred outperforms the existing sigma70 promoter prediction approaches in terms of accuracy and stability. Additionally, this approach can also be extended to predict promoters of other species. In order to facilitate experimental biologists, an online web server for the proposed method was established, which is freely available at http://server.malab.cn/70ProPred/.

Keywords: sigma70 promoter, PSTNP_{SS}, PseEIIP, SVM

Background

Transcription is strictly regulated and controlled by cis-regulatory DNA elements, which were known as promoters and enhancers. These elements control the level of gene expression and cell fate. Promoters are cis-acting DNA sequences that switch on or off the gene expression. They are generally located upstream of the transcription start sites of genes. In prokaryotes, promoters are identified by RNA polymerase and a related sigma factor [1]. Different σ factors interact with well-defined consensus promoter sequences. Each σ factor is marked according to its molecular weight. σ^{70} is a well-known factor that regulates the transcription of most house-keeping genes in normal circumstances [1]. For σ^{70} promoter, two short sequence elements approximately located at around -10 bp and -35 bp nucleotides upstream from the transcription start site (TSS) with consensus TATAAT and TTGACA respectively [2]. It is important to identify the promoters in a genome, because it can help clarify the regulatory mechanism in the genome and explain disease-causing variants within cis-regulatory elements [3, 4]. Meanwhile, it's a crucial part of gene structure recognition and the core issue of building gene transcriptional regulation network. Man's understanding of promoter is developing all the time. It's an area of great concern as people place increasing

* Correspondence: zouquan@nclab.net
[1] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
Full list of author information is available at the end of the article

attention on their importance not only in developmental gene expression but also in environmental response [5, 6].

Due to the rapid development of genome sequencing technology, large-scale data has been generated [7–9], the stable and accurate identification of promoter is an important problem. Because standard laboratory methods are time-consuming and performance overhead costing, bioinformatics technologies with perfect precision represent the ideal alternative for massive fast recognition of promoter.

The σ^{70} promoter recognition task is a binary classification task. Feature extraction and classifier design are the key problems in promoter identification technology. In the past 20 years, based on the feature of promoter sequences, a serious of approaches have been developed for detecting promoter region in prokaryotes [10–23]. In 2007, Zhang proposed an algorithm using increment of diversity with quadratic discriminant (IDQD) analysis [21]. Position weight matrix (PWM) is always regarded as a description of the sequence information, but sometimes it gives poor results [22]. After that, Wu proposed an improved Position Weight Matrix (IPWM) [23] in 2011. In the same year, Lin proposed a hybrid method (IPMD), which combines location-related scoring function and diversity increment with improved Mahalanobis Discriminant to predict promoters [16]. The next year, variable-window Z-curve was used for extracting basic features of prokaryotic promoter [14]. Recently, Lin developed an improved Z-curve called 'multi-window Z-curve' (PseZNC), which can express the frequency characteristics and three dimensionality characteristics of different length sequences [15], etc. They were mainly focused on the σ^{70} promoter recognition. Among these approaches, some typical machine learning algorithms have been used to develop prokaryotic promoter region prediction. Such as, SVM (Support Vector Machine), RF (Random Forest), NB (Naïve Bayes), PLS (Partial Least Square), etc. Although these approaches have contributed to the advancement of promoter recognition, their performance demonstrates that there is a long way to go to predict promoter accurately for the following reasons. (i) Most existing approaches overlooked the correlation of neighboring nucleotides in each position, especially their difference in positive and negative samples. (ii) The local biological and physical properties of DNA may have a certain relationship with the promoters, which plays an important role in identifying them but were utterly ignored. (iii) Few web-servers were provided as the predictors, and hence their usage is quite limited [24].

In current research, we exploited a new bioinformatics tool called 70ProPred, to predict σ^{70} promoter through a combination of position-specific trinucleotide propensity (PSTNP) and electron-ion interaction pseudopotentials (EIIPs) of nucleotides. Finally, based on the results analysis of jackknife test, 70ProPred significantly outperforms existing prediction models, and should be useful for identifying σ^{70}promoter.

Methods

Briefly, 70ProPred is a prediction model based on support vector machine (SVM), which was built by PSTNP_{SS} and PseEIIP sequence coding strategies. An outline of the computational framework of 70ProPred predictor is shown in Fig. 1.

Datasets

In the present study, we used the latest datasets in both [15] and [16]. A total of 741 σ^{70} promoter samples were selected from the *E.coli* K-12 genome, which have been verified by experiments and downloaded from the RegulonDB9.0 (http://regulondb.ccg.unam.mx/) [25]. The promoter region [TSS-60…TSS + 20] (the locus of TSS is 0) were prepared as positive samples with a length of 81 bp.

As there are not enough experimental confirmed negative sequences, negative samples are collected from both coding and non-coding regions. In simple terms, the benchmark dataset S used in this study can be expressed as:

$$\begin{cases} S = S^{+} \cup S^{-} \\ S^{-} = S^{-}_{\text{coding}} \cup S^{-}_{\text{non-coding}} \end{cases} \tag{1}$$

Where S^{+} contains 741 σ^{70} promoter samples, S^{-} contains 1400 non-promoter samples, S^{-}_{coding} contains 700 coding sequences, $S^{-}_{\text{non-coding}}$ contains 700 convergent intergenic sequences. Each sample contains 81 nucleotides, which is selected by a sliding window. Additionally, symbol ∪ means union.

Analysis of σ^{70} promoter samples for conserved motif composition

The MEME Suite is designed to screen common sequence motifs from a set of sequences (amino acid or nucleotide). A motif can be assumed to be a conservative sequence pattern that repeats itself over a set of related sequences [26]. MEME is a useful sequence analysis tool that can rapidly detect new, non-gapped motifs for biological sequence data (protein, DNA and RNA) [27]. Then, we applied this tool to analyze the main motifs of σ^{70} promoter samples and found that only a small part of these samples which have corresponding motifs in Fig. 2. The maximum number of motif was set to 3 and the remaining arguments were set to default.

Although phylogenetic foot printing takes the advantage of relatively conservative of motifs between species [28], these motifs are short and not complete species [29, 30]. For example, in Table 1, the number of sites

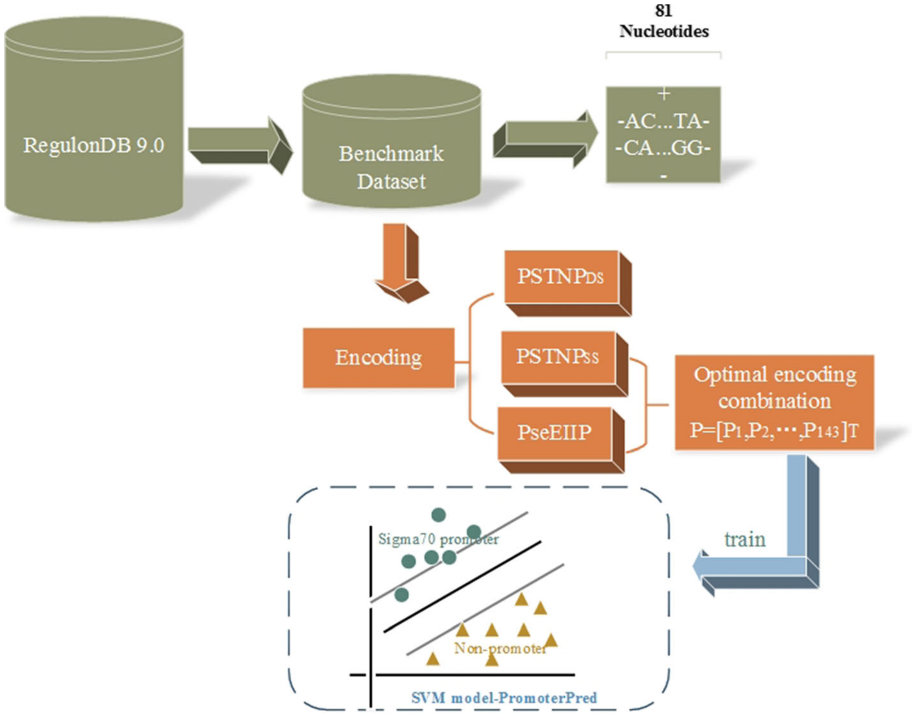


Fig. 1 Overview of the proposed 70ProPred predictor. The diagram mainly contains datasets, sequence descriptors and 70ProPred prediction system. The optimal encoding combination PSTNP$_{SS}$ and PseEIIP are used as the input to train a SVM classifier. After optimization of the SVM parameters, the best SVM model is constructed based on the jackknife performance

contributing to the construction of motif 1 only 47, which may result in a great deal of false positive results. Therefore, it would be practicable to turn to the machine learning-based methods and has been proved to be effective in many fields [26, 31–46].

Feature vector construction

In order to get more useful information from the sequence, we extracted two kinds of features. Position-specific tendencies of trinucleotide (PSTNPss or PSTNP$_{DS}$) was adopted to reveal the differences in the

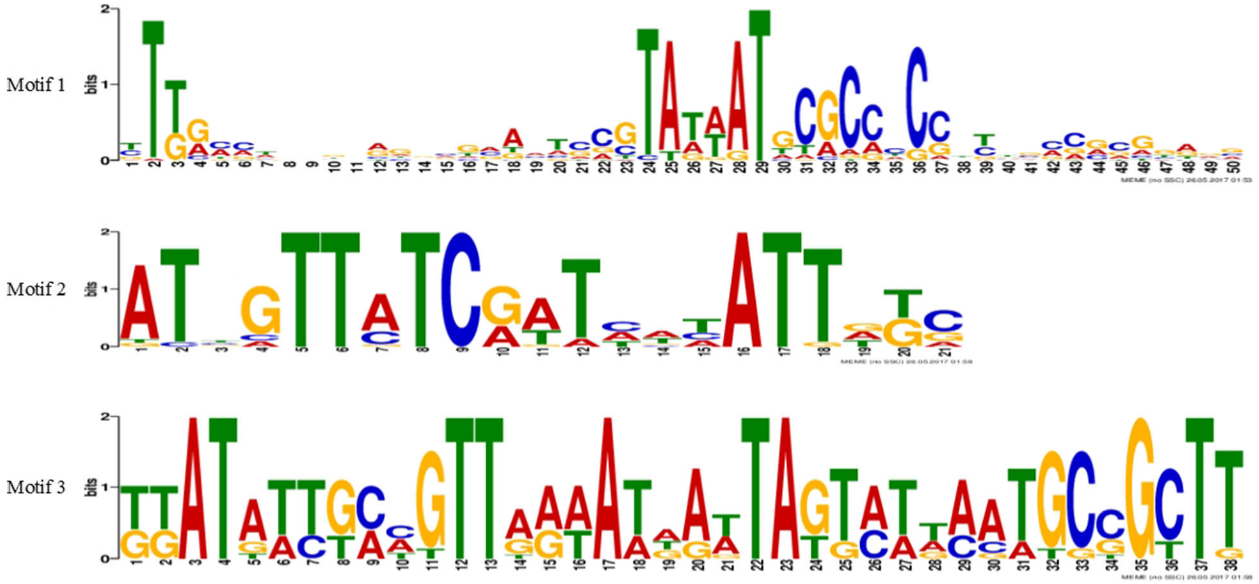


Fig. 2 Motif of σ^{70} promoter samples as found by the MEME system. The corresponding three-motif logos as visualized for σ^{70} promoter samples (details in Table 1)

Table 1 Conserved motifs of σ^{70} promoter samples identified by the MEME system

Motif	Width	Best possible match	Sites count
1	50	YTKRMMWNNBNRGNVGVAMTSCGTATWATGCGCCYCCNYBVMCVCGKRVV	47
2	21	ATBGTTATCRATHWHATTDKC	20
3	38	KKATATTGMHGTTRRWATDAWTAGTMTWAATGCSGCTT	10

distribution of all positive and negative samples between trinucleotide. While the electron-ion interaction pseudopotentials (PseEIIP) were adopted to represent the interaction of trinucleotides.

Position-specific trinucleotide propensity based on single-stranded or double-stranded characteristic of DNA (PSTNP$_{SS}$/PSTNP$_{DS}$)

The recently proposed position-specific theory has been successfully applied to numerous fields of bioinformatics, such as identifying protein S-nitrosylation sites [47], hydroxyproline and hydroxylysine sites [48], DNA replication origin sites [49], Enhancer [41], etc. Besides, a series of studies have shown that the trinucleotides composition was effective in recognition of promoter [28, 50–52]. Inspired by the above studies, we presented a novel feature extraction strategy, which combined position-specific trinucleotide property (PSTNP) with the K-mer nucleotide composition information to predict σ^{70} promoter. A brief account of feature extraction is as follows.

Let S represents a sample which was consisted of A, G, C, and T, i.e.

$$S = N_1 N_2 N_3 \cdots N_L \tag{2}$$

In which L means the length of the sample and $L = 81$, and

$$N_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} (i = 1, 2, \ldots, L) \tag{3}$$

represents the i-th position of corresponding nucleotide in the sequence.

PSTNP$_{SS}$

Feature PSTNP$_{SS}$ using a statistical strategy based on single-stranded characteristics of DNA. There are $4^3 =$ 64trinucleotides: AAA, AAC, AAG, ..., TTT. So, for an 81 bp sample, its details of the trinucleotides position specificity can be expressed by the following 64 × 79 matrix [41]:

$$Z = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,79} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,79} \\ \vdots & \vdots & \cdots & \vdots \\ z_{64,1} & z_{64,2} & \cdots & z_{64,79} \end{bmatrix} \tag{4}$$

where the variable

$$z_{i,j} = F^{+}(3mer_i|j) - F^{-}(3mer_i|j) \quad (i = 1, 2, \ldots, 64; j = 1, 2, \ldots 79) \tag{5}$$

$F^{+}(3mer_i|j)$ and $F^{-}(3mer_i|j)$ denote the frequency of the i-th trinucleotide (3mer$_i$) at the j-th position appear in the positive (S^{+}) and negative (S^{-}) data sets, respectively. In the formula, 3mer$_1$ equals AAA,3mer$_2$ equals AAC, ..., 3mer$_{64}$ equals TTT.

Therefore, the sample of Eq.2 can be expressed as:

$$\mathrm{S} = [\phi_1, \phi_2, \ldots, \phi_u, \ldots, \phi_{79}]^T \tag{6}$$

where T is the operator of transpose and ϕ_u was defined as follows:

$$\phi_u = \begin{cases} z_{1,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{AAA} \\ z_{2,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{AAC} \\ z_{3,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{AAG} \\ \vdots & \quad \vdots \qquad \vdots \\ z_{64,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{TTT} \end{cases} \quad (1 \le u \le 79) \tag{7}$$

PSTNP$_{DS}$

Feature PSTNP$_{DS}$ using a statistical strategy based on double-stranded characteristics of DNA according to complementary base pairing, so they have more evident statistical features. At this point, we deem A and T as identical, the same to C and G. Thus, for every sample, it can be converted into a sequence contained A and T only. As shown in Fig. 3, promoter-1 converted into promoter-1 AC.

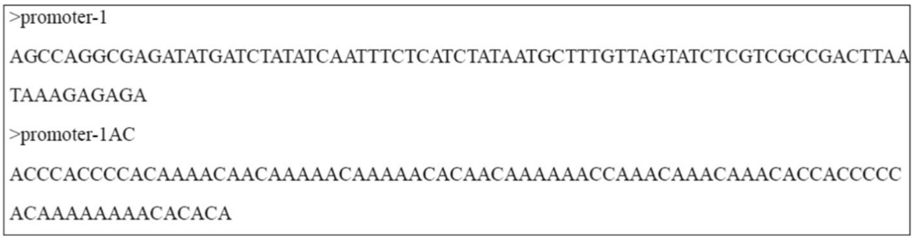
So, there are $2^3 = 8$ trinucleotides: AAA, AAC, ..., CCC. Therefore, for an 81 bp sample, its details of the trinucleotides position specificity can be expressed by the following 8 × 79 matrix:

$$Z' = \begin{bmatrix} z'_{1,1} & z'_{1,2} & \cdots & z'_{1,79} \\ z'_{2,1} & z'_{2,2} & \cdots & z'_{2,79} \\ \vdots & \vdots & \cdots & \vdots \\ z'_{8,1} & z'_{8,2} & \cdots & z'_{8,79} \end{bmatrix} \tag{8}$$

where the variable

$$z'_{i,j} = F^{+}(3mer_i|j) - F^{-}(3mer_i|j) \quad (i = 1, 2, \ldots, 8; j = 1, 2, \ldots 79) \tag{9}$$

$F^{+}(3mer_i|j)$ and $F^{-}(3mer_i|j)$ denote the frequency of the i-th trinucleotide (3mer$_i$) at the j-th position appear in the positive (S^{+}) and negative (S^{-}) data sets, respectively.



>promoter-1

AGCCAGGCGAGATATGATCTATATCAATTTCTCATCTATAATGCTTTGTTAGTATCTCGTCGCCGACTTAA

TAAAGAGAGA

>promoter-1AC

ACCCACCCCACAAAACAACAAAAACAAAAACACAACAAAAAACCAAACAAACAAACACCACCCCC

ACAAAAAAAACACACA

Fig. 3 promoter-1 converted into promoter-1 AC

In the formula, 3mer_1 equals AAA,3mer_2 equals AAC, ..., 3mer_8 equals CCC.

Therefore, the sample of Eq.2 can be expressed as:

$$\mathrm{S}' = \left[\phi_1', \phi_2', \dots, \phi_u', \dots, \phi_{79}'\right]^T \tag{10}$$

where T is the operator of transpose and ϕ_u' was defined as follows:

$$\phi'_u = \begin{cases} z'_{1,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{AAA} \\ z'_{2,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{AAC} \\ z'_{3,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{ACA} \\ \vdots & \quad \vdots \qquad\qquad \vdots \\ z'_{8,u}, & \text{when } \mathrm{N}_u\mathrm{N}_{u+1}\mathrm{N}_{u+2} = \mathrm{CCC} \end{cases} \quad (1 \le u \le 79) \tag{11}$$

Electron-ion interaction pseudopotentials of trinucleotide (PseEIIP)

Nair [53] came up with electron-ion interaction pseudopotentials (EIIP) value of nucleotides A, G, C, T. The EIIP value based methods have been shown effective through previous studies, such as the recognition of gene F56F11.4, prediction of the cystic-fibrosis gene [54], recognition of enhancer [41], and so on [55, 56].

The electron-ion interaction pseudopotentials value for the nucleotides [53] are shown in Table 2. We let $EIIP_A$, $EIIP_T$, $EIIP_G$, and $EIIP_C$ denote the EIIP values of nucleotides A, T, G and C, respectively. Then, we employed the mean EIIP value of trinucleotides in each sample to construct feature vector, which can be formulated as:

$$V = [EIIP_{AAA} \cdot f_{AAA}, EIIP_{AAC} \cdot f_{AAC}, \dots, EIIP_{TTT} \cdot f_{TTT}] \tag{12}$$

Table 2 EIIP values of nucleotides

Nucleotide	EIIP(Ry)
A	0.1260
T	0.1335
G	0.0806
C	0.1340

where f_{xyz} equal to the normalized frequency of the i-th trinucleotide (3mer_i), $EIIP_{xyz} = EIIP_x + EIIP_y + EIIP_z$ expresses the EIIP value of one trinucleotide and X, Y, Z∈ [A, C, G, T]. Obviously, the dimension of vector V is 64.

Model building and parameter selection

SVM classification algorithm plays a significant role in some areas of bioinformatics [18, 36, 40, 57]. In this work, SVM was implemented using the LIBSVM packet [58] to build models and execute predictions. The radial basis function (RBF) was selected as the kernel function. At the same time, penalty parameter C and kernel parameter γ were optimized using SVMcg in the LIBSVM package. The final parameters C = 22.6274 and γ = 2.8284 were selected for the prediction of σ^{70} promoters and non-promoters.

The jackknife test is regarded as a unique random test that can produce the unique result for a given dataset [59]. Therefore, all these parameters were optimized through jackknife test.

In order to evaluate the predictive performance of the model, four metrics are calculated: sensitivity (Sn), specificity (Sp), accuracy (Acc) and MCC:

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$SP = \frac{TN}{TN + FP} \tag{14}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

In the formula, TP, TN, FP and FN represent the numbers of true positives (accurately predicted σ^{70} promoters), true negatives (accurately predicted non-promoters), false positives (falsely predicted σ^{70} promoters) and false negatives (falsely predicted non-promoters).

Results and discussion

Prediction of σ^{70} promoter using only PSTNP

PSTNP was first proposed for predicting enhancer [41]. The obvious advantage of this approach is that the feature vectors are encoded in a way that contains information from all training samples. In this work, the ability of $PSTNP_{SS}$ and $PSTNP_{DS}$ to discriminate σ^{70} promoter and non-promoter were first declared by jackknife test (Table 3). For σ^{70} promoter, the $PSTNP_{DS}$ model obtained a good performance, reaching at 75.98% sensitivity, 88.57% specificity, 84.21% accuracy and the 0.6493 of MCC value, while the $PSTNP_{SS}$ model obtained a Sn of 90.82%, a Sp of 96.57%, an Acc of 94.58% and a MCC of 0.8797.

A comparative figure (Fig. 4) with *F-value* of trinucleotides in different position also declared the difference in forecast results.

Furthermore, we used the Entropy (Additional file 1) [23, 60] to demonstrate the conservative sites of trinucleotides property in σ^{70} promoter. In order to comparing, the calculated entropy values for the trinucleotide of the σ^{70} promoter and non-promoter are shown in Fig. 5. Obviously, the lower entropy, the more conservative the position is. From this figure, we can see that in σ^{70} promoter most sites are obviously of lower entropy values compared with non-promoter. It may prove that $PSTNP_{SS}$ feature extraction agreed well with the previous prediction that in different positions the trinucleotide is conservative in σ^{70}promoter.

From all above, the composition of trinucleotide which based on single-stranded characteristics of DNA contributes to the detection of σ^{70} promoter. This result indicated that the $PSTNP_{SS}$ model performs better than the $PSTNP_{DS}$ model in σ^{70} promoter prediction. Consequently, the training model optimized on the basis of the feature $PSTNP_{SS}$.

Improving performance by incorporating PseEIIP

Because the physicochemical property indexes of nucleotides affect the recognition of promoter, incorporating the sample's average energy of delocalized electrons (EIIP), especially, the EIIP value of trinucleotides with $PSTNP_{SS}$ might boost the performance of the training model, the prediction results are listed in Table 4.

Eventually, the prediction model was established using the $PSTNP_{SS}$ + PseEIIP feature extraction methods combined with the SVM classifier (cost parameter −c 22.6274, −g 2.8284) to predict σ^{70}promoter.

Table 3 Jackknife test performance of $PSTNP_{SS}$ and $PSTNP_{DS}$

Features	Sn (%)	Sp (%)	Acc (%)	MCC	SVM
$PSTNP_{SS}$ (79)	90.82	96.57	94.58	0.8797	-c 22.6274 -g 1.4142
$PSTNP_{DS}$ (79)	75.98	88.57	84.21	0.6493	-c 1.4142 -g 2.8284 -w1 1.2 -w-1 1

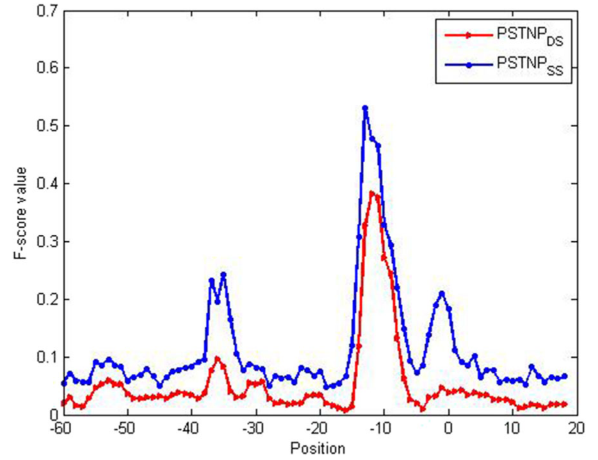


Fig. 4 *F-score* value of trinucleotides in both $PSTNP_{SS}$ and $PSTNP_{DS}$

In order to gauge the predictive performance of training model, the ROC curve and the area under the ROC curve (AUC) were adopted. The AUC value the 70ProPred model was 0.990 (Fig. 6).

Furthermore, we constructed a heat map to visually show the distribution of feature PseEIIP in positive dataset, as shown in Fig. 7. Each hotspot in the heat map corresponds to a unique trinucleotide; for instance, hotspot (1, 1) corresponds to triplet AAA. For more detailed information on the heat map, please see Additional file 2: Table S1. Red squares are positively associated with recognition ability.

Comparison of the performance between SVM and other classifiers

In order to select a better classifier to identify σ^{70} promoter, we analyzed the performances of KNN [61],

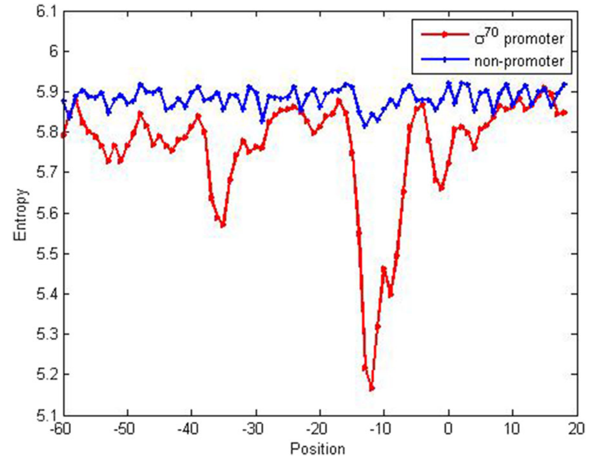


Fig. 5 Entropy of trinucleotide in the σ^{70} promoter and non-promoter

Table 4 Performances of our model on the jackknife test

Features	Sn (%)	Sp (%)	Acc (%)	MCC
PSTNP_{SS} (79)	90.82	96.57	94.58	0.8797
PSTNP_{SS} (79) + PseEIIP (64)	93.12	96.86	95.56	0.9018

NB [62], RF [63], EB [64], LibD3C [65], GBDT [66] and SVM, which are the most widely used classification algorithms in bioinformatics. Since it is known that the number of neighbors has an impact on the performance of KNN algorithm and the tree number has an impact on the performance of RF algorithm, the optimal parameters of KNN and RF are searched in the study, as shown in Additional file 3: Table S2 and Additional file 4: Table S3.

The performances for the above classifiers in the jackknife test are shown in Table 5. The results indicate that SVM performs better than the other classifiers with the highest MCC value 0.9018.

Comparison of 70ProPred with other σ^{70} promoter prediction methods

The 5-fold cross-validation results achieved by Z-curve [14], PseZNC [15], IPMD [16], PSTNP_{DS} and 70ProPred on the benchmark dataset are listed in Table 6. Their marked difference is the feature extraction strategy. Therefore, the advantage of 70ProPred superior to other methods is mainly due to the combination of PSTNP_{SS} and PseEIIP coding strategy. PSTNP_{SS} employs primary sequence information of trinucleotides, and PseEIIP is closely related to the physicalchemical property of DNA sample. As shown in Table 3, based on the PSTNP_{SS} feature only, the prediction result of 70ProPred was significantly higher than the other methods. The performance has been further improved by adding the feature PseEIIP, as shown in Table 4, especially in sensitivity (Sn). Taken together, the application of feature PSTNP_{SS} and feature PseEIIP achieved a good performance in the prediction of σ^{70}promoter.

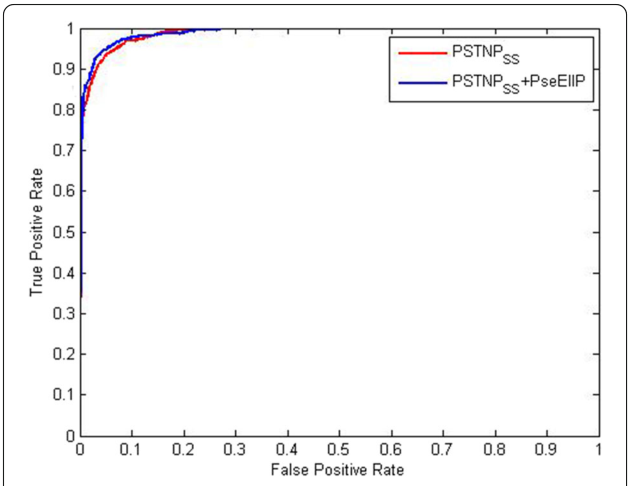


Fig. 6 The ROC curves to assess the predictive performance based on different sequences encoding schemes for σ^{70}promoter

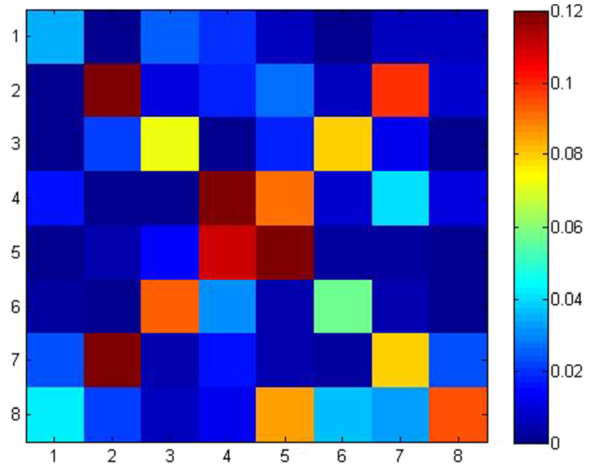


Fig. 7 A heat map for the *F-score* values of the 64 trinucleotides with different EIIP values. The blue boxes indicate the features with a lower effect for recognition of the σ^{70} promoter, while the red boxes indicate the features that are useful for the recognition of the σ^{70}promoter

The results in Table 6 also show that the PSTNP_{DS}-based model performs better than the multi-window Z-curve-based method. It can also be a supplement to the present methods for predicting other DNA related predictions.

Conclusions

The 70ProPred is a new bioinformatics tool for predicting σ^{70} promoter. This tool uses the feature extraction methods of PSTNP_{SS} and PseEIIP. The combination of features and SVM could achieve an overall MCC value of 0.90. Compared to other σ^{70} promoter prediction models, 70ProPred produced better results. Although this method shows good performance in σ^{70} promoter prediction, there is still room to improve prediction performance due to the following reasons. (i) Since structural information is a supplementary to sequence

Table 5 Comparison of different classifiers for identifying σ^{70}promoter

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
KNN (8)	87.04	96.21	93.04	0.8450
Naïve Bayes	91.90	89.00	90.00.	0.7891
Random Forest (200)	85.29	97.79	93.46	0.8548
Ensembles for Boosting (200)	89.88	95.29	93.41	0.8541
LibD3C	77.33	87.57	84.03	0.6478
GBDT	86.50	96.14	92.81	0.8397
SVM	93.12	96.86	95.56	0.9018

Table 6 performances of our model, Z-curve, $PSTNP_{DS}$, PseZNC and IPMD on 5-fold cross-validation

Methods	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
Z-curve	74.6	79.5	77.8	0.527	0.848
$PSTNP_{DS}$	75.9	88.0	83.8	0.641	0.911
PseZNC	80.3	86.8	84.5	0.663	0.909
IPMD	82.4	90.7	87.9	0.731	–
70ProPred	92.4	96.9	95.3	0.897	0.990

information, the future work may build a model combine with the two aspects. (ii) The feature selection algorithms can be used to delete the redundant features to improve the prediction model. (iii) More species of promoters should be adopted to estimate the performance of 70ProPred method. In conclusion, our future work is to extend this method to other species promoter region prediction. We suspect that our feature extraction methods is not only suitable for identifying promoter, but also for other bioinformatics sequence classification tasks.

Availability

The web-server for 70ProPred has been established. It is now freely available to all interested users at http://server.malab.cn/70ProPred/. All the data sets used in this study can also be download on the website.

Additional files

Additional file 1: Entropy. (DOC 94 kb)

Additional file 2: **Table S1.** Rules of composition of heat maps. (DOC 41 kb)

Additional file 3: **Table S2.** Comparison prediction results of different k neighbors. (DOC 47 kb)

Additional file 4: **Table S3.** Comparison prediction results of different nTrees. (DOC 42 kb)

Abbreviations

AUC: Area Under the ROC Curve; EB: Ensembles for Boosting; GBDT: Gradient Boosting Decision Tree; KNN: k-Nearest Neighbor; MCC: Matthew's Correlation Coefficient; NB: Naïve Bayes; PseEIIP: Electron-ion interaction pseudopotentials of nucleotides; $PSTNP_{DS}$: Position-specific trinuclotide propensity based on double-stranded characteristic of DNA; PSTNPss: Position-specific trinuclotide propensity based on single-stranded characteristic of DNA; RF: Random Forest

Acknowledgements

The work was supported by the Natural Science Foundation of China (No. 61771331).

Funding

Publication costs were funded by the Natural Science Foundation of China (No. 61771331).

Availability of data and materials

All data analyzed during this study are included in this published article.

About this supplement

This article has been published as part of *BMC Systems Biology* Volume 12 Supplement 4, 2018: Selected papers from the 11th International Conference on Systems Biology (ISB 2017). The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-12-supplement-4.

Authors' contributions

QZ conceived the research. WH performed the study. CJ, QZ and YD provided valuable suggestions and improvements. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China. [2]Department of Mathematics, Dalian Maritime University, Dalian 116026, China. [3]College of Information and Technology, Hainan University, Haikou 570228, China.

Published: 24 April 2018

References

1. Barrios H, Valderrama B, Morett E. Compilation and analysis of sigma(54)-dependent promoter sequences. Nucleic Acids Res. 1999;27(22):4305–13.
2. Hawley DK, McClure WR. Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res. 1983;11(8):2237–55.
3. Li Y, Chen CY, Wasserman WW. Deep feature selection: theory and application to identify enhancers and promoters. J Comput Biol. 2016;23(5): 322–36.
4. Geng Y, Zhao ZM, Zhang XP, Wang WK, Cui XJ, Ye K, et al. An improved burden-test pipeline for identifying associations from rare germline and somatic variants. BMC Genomics. 2017;18(7):55–62.
5. Bernardo LM, Johansson LU, Skarfstad E, Shingler V. sigma54-promoter discrimination and regulation by ppGpp and DksA. J Biol Chem. 2009;284(2): 828–38.
6. Lu C, Xie MC, Wendl MC, Wang JY, McLellan MD, Leiserson MDM, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun. 2015;6
7. Mrozek D, Malysiak-Mrozek B, Klapcinski A. Cloud4Psi: cloud computing for 3D protein structure similarity searching. Bioinformatics. 2014;30(19):2822–5.
8. Mrozek D, Danilowicz P, Malysiak-Mrozek B. HDInsight4PSi: boosting performance of 3D protein structure similarity searching with HDInsight clusters in Microsoft azure cloud. Inform Sciences. 2016;349:77–101.
9. Mrozek D, Gosk P, Malysiak-Mrozek B. Scaling ab initio predictions of 3D protein structures in Microsoft azure cloud. J Grid Comput. 2015;13(4):561–85.
10. Zhou X, Li Z, Dai Z, Zou X. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform. J Theor Biol. 2013;319: 1–7.
11. Wang H, Benham CJ. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC Bioinformatics. 2006;7:248.
12. Towsey MW, Gordon JJ, Hogan JM. The prediction of bacterial transcription start sites using SVMs. Int J Neural Syst. 2006;16(5):363–70.
13. Towsey M, Hogan JM, Mathews S, Timms P. The in silico prediction of promoters in bacterial genomes. Genome Inform. 2007;19:178–89.
14. Song K. Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. Nucleic Acids Res. 2012;40(3):963–71.
15. Lin H, Liang ZY, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans Comput Biol Bioinform. 2017;PP(99):1–1.
16. Lin H, Li QZ. Eukaryotic and prokaryotic promoter prediction using hybrid approach. Theory Biosci. 2011;130(2):91–100.

17. Kanhere A, Bansal M. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics. 2005;6:1.
18. Jia CZ, He WY, Yao YH. OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. J Biomol Struct Dyn. 2017;35(4):829–35.
19. Huerta AM, Collado-Vides J. Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol. 2003;333(2):261–78.
20. de Avila ESS, Gerhardt GJ, Echeverrigaray S. Rules extraction from neural networks applied to the prediction and recognition of prokaryotic promoters. Genet Mol Biol. 2011;34(2):353–60.
21. ZHANG Y, JIA Y, Lü J. THE RECOGNITION OF σ~(70) PROMOTERS IN Escherichia coli K-12. Acta Biophysica Sinica. 2007;6:011.
22. Barrick D, Villanueba K, Childs J, Kalil R, Schneider TD, Lawrence CE, et al. Quantitative analysis of ribosome binding sites in E.Coli. Nucleic Acids Res. 1994;22(7):1287–95.
23. Wu Q, Wang J, Yan H. An improved position weight matrix method based on an entropy measure for the recognition of prokaryotic promoters. Int J Data Min Bioinform. 2011;5(1):22–37.
24. Mrozek D, Malysiak-Mrozek B, Siaznik A. Search GenBank: interactive orchestration and ad-hoc choreography of web services in the exploration of the biomedical resources of the National Center for Biotechnology Information. Bmc Bioinformatics. 2013;14(1):1–19.
25. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44(D1):D133–43.
26. Liao Z, Wang X, Zeng Y, Zou Q. Identification of DEP domain-containing proteins by a machine learning method and experimental analysis of their expression in human HCC tissues. Sci Rep. 2016;6:39655.
27. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39–49.
28. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014;42(21): 12961–72.
29. Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y. Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res. 2008; 18(2):310–23.
30. Abeel T, Saeys Y, Rouze P, Van de Peer Y. ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics. 2008;24(13):i24–31.
31. Liao Z, Wang X. Chen X. Comb Chem High Throughput Screen: Zou Q. Prediction and Identification of Kruppel-like transcription factors by machine learning method; 2017.
32. Xiong H, Liu D, Li Q, Lei M, Xu L, Wu L, et al. RED-ML: a novel, effective RNA editing detection method based on machine learning. Gigascience. 2017; 6(5):1–8.
33. Chen W, Xing P, Zou Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. Sci Rep. 2017;7: 40242.
34. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 2017;8(3): 4208–17.
35. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep. 2017;7:42362.
36. Jia CZ, Liu T, Wang ZP. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. Mol BioSyst. 2013;9(11):2909–13.
37. Jia C, Lin X, Wang Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. Int J Mol Sci. 2014;15(6):10410–23.
38. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013;41(6): e68.
39. Kabir M, Hayat M. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Mol Gen Genomics. 2016;291(1):285–96.
40. Jia C, He W. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. Sci Rep. 2016;6:38741.
41. He W, Jia C. EnhancerPred2.0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron-ion interaction potential feature selection. Mol BioSyst. 2017;13(4):767–74.
42. Liu B, Fang L, Wang S, Wang X, Li H, Chou KC. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J Theor Biol. 2015; 385:153–9.
43. Liu B, Fang L, Long R, Lan X, Chou KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016;32(3):362–9.
44. Chen W, Feng PM, Deng EZ, Lin H, Chou KC. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem. 2014;462:76–83.
45. Liu Z, Xiao X, Qiu WR, Chou KC. iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem. 2015;474:69–77.
46. Jia J, Zhang L, Liu Z, Xiao X, Chou KC. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics. 2016;32(20): 3133–41.
47. Xu Y, Shao XJ, Wu LY, Deng NY, Chou KC. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013;1:e171.
48. Xu Y, Wen X, Shao XJ, Deng NY, Chou KC. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014;15(5):7594–610.
49. Xiao X, Ye HX, Liu Z, Jia JH, Chou KC. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget. 2016; 7(23):34180–9.
50. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273(1):236–47.
51. Chen W, Lin H, Chou K-C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015; 11(10):2620–34.
52. Pei Li MG. Chunyu Wang, Xiaoyan Liu, Quan Zou. An overview of SNP interactions in genome-wide association studies. Briefings in Functional Genomics. 2015;14(2):143–55.
53. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation. 2006;1(6):197–202.
54. Rao KD, Swamy M. Analysis of genomics and proteomics using DSP techniques. IEEE Transactions on Circuits and Systems I: Regular Papers. 2008;55(1):370–8.
55. Sahu SS, Panda G. Efficient localization of hot spots in proteins using a novel S-transform based filtering approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2011;8(5):1235–46.
56. Khan A. G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties. Anal Biochem. 2011;412(2):173–82.
57. Wei L, Liao M, Gao X, Zou Q. Enhanced protein fold prediction method through a novel feature extraction technique. IEEE Transactions on Nanobioscience. 2015;14(6):649–59.
58. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27.
59. Chou K-C, Shen H-B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science. 2010;2(10):1090.
60. Fickett JW. Quantitative discrimination of MEF2 sites. Mol Cell Biol. 1996; 16(1):437–41.
61. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.
62. Rish I. An empirical study of the naive Bayes classifier. Journal of Universal Computer Science. 2001;1(2):127.
63. Ho TK. Random decision forests. International Conference on Document Analysis and Recognition. IEEE Computer Society. 1995;278.
64. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res. 1999;11:169–98.
65. Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zou Q. LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing. 2014; 123:424–35.
66. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.