Identifying Sigma70 Promoters with Novel Pseudo Nucleotide Composition

Hao Lin, Zhi-Yong Liang, Hua Tang , and Wei Chen

Abstract—Promoters are DNA regulatory elements located directly upstream or at the 5' end of the transcription initiation site (TSS), which are in charge of gene transcription initiation. With the completion of a large number of microorganism genomics, it is urgent to predict promoters accurately in bacteria by using the computational method. In this work, a sequence-based predictor named "iPro70-PseZNC" was designed for identifying sigma70 promoters in prokaryote. In the predictor, the samples of DNA sequences are formulated by a novel pseudo nucleotide composition, called PseZNC, into which the multi-window Z-curve composition and six local DNA structural properties are incorporated. In the 5-fold cross-validation, the area under the curve of receiver operating characteristic of 0.909 was obtained on our benchmark dataset, indicating that the proposed predictor is promising and will provide an important guide in this area. Further studies showed that the performance of PseZNC is better than it of multi-window Z-curve composition. For the sake of convenience for researchers, a user-friendly online service was established and can be freely accessible at http://lin.uestc.edu.cn/server/iPro70-PseZNC. The PseZNC approach can be also extended to other DNA-related problems.

Index Terms—Prokaryote, sigma70 promoter, PseZNC, multi-window Z-curve, local DNA structural property

1 Introduction

The promoter is a region of DNA element that located near the transcription start sites (TSS) and initiates the transcription of a gene. In bacteria, the $\boldsymbol{\sigma}$ (sigma) factor of RNA holoenzyme plays a critical role in recognizing and binding to the promoter sequence during gene transcription. Diferent $\boldsymbol{\sigma}$ factors can recognize different promoter sequences. Thus, the type of bacterial promoters is defined by the type of $\boldsymbol{\sigma}$ factor. Thereinto, the sigma70 factor ($\boldsymbol{\sigma}^{70}$), also called the primary sigma factor or "housekeeping" sigma factor, regulates the transcription of the most genes in growing cells [1], [2]. Therefore, for revealing the mechanism of transcription of most of genes, further research on the $\boldsymbol{\sigma}^{70}$ promoter is crucial, which will help us to understand the subsequent steps of gene expression and establish the network of gene transcription.

The correct identification of $\boldsymbol{\sigma}^{70}$ promoter is the first step for understanding its regulatory mechanisms. Though more details about $\boldsymbol{\sigma}^{70}$ promoters can be obtained by using biochemical experimental technique, the wet-experimental approaches are time-consuming and expensive. With the development of high-throughput sequencing technologies and information technologies, a variety of DNA sequences are generated, which makes it possible to identify $\boldsymbol{\sigma}^{70}$ promoters by computational methods. Owing to the advantages of computational methods, some algorithms such as increment of diversity with quadratic discriminant (IDQD) [3], artificial neural network (ANN) [4], hidden Markov models (HMMs) [5], support vector machine (SVM) [6], position weight matrix [7], and so on, have been developed to identify $\boldsymbol{\sigma}^{70}$ promoters. These methods have yielded encouraging results, and each of them did play a role in the field of $\boldsymbol{\sigma}^{70}$ promoter prediction, however, further study on $\boldsymbol{\sigma}^{70}$ promoter prediction is needed due to the following reasons. (i) The datasets used to construct the prediction models in these methods were not enough to reflect the common features of promoters in statistics. (ii) Most existing approaches only took the short or local DNA sequence information into account, and ignored the facts that the long or global DNA sequence information are important for the prediction of promoters. (iii) No web-server was provided as the predictor, thus resulting in a lack of practicality for other related researchers.

Thus, in the present study, a novel method is proposed to improve $\boldsymbol{\sigma}^{70}$ promoter prediction from the above three aspects.

2 Materials and Methods

2.1 Benchmark Dataset

In this work, a high quality benchmark dataset was constructed to train and test our method. Total of 741 $\boldsymbol{\sigma}^{70}$ promoter sequences were chosen from the genome of *E. coli* K-12, which were confirmed by experimental technologies and were dowloaded from the RegulonDB 9.0 (http://regulondb.ccg.unam.mx/) [8]. All of the positive samples have the length of 81bp, i.e. from -60 to +20 bp with the TSS in their fragments (TSSs are at the 0_{th} site).

- *H. Lin and Z.Y. Liang are with the Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China. E-mail: hlin@uestc.edu.cn, liangzhiyong07140@gmail.com.*
- *H. Tang is with the Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China. E-mail: Tanghua771211@aliyun.com.*
- *W. Chen is with the Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China. E-mail: greatchen@ncst.edu.cn.*

Manuscript received 12 Jan. 2017; accepted 2 Feb. 2017. Date of publication 8 Feb. 2017; date of current version 5 Aug. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2666141

TABLE 1
The Positive and Negative Samples

Data set	Positive samples	Negative samples
Sample source	741 $\boldsymbol{\sigma}^{70}$ promoters of *E. coli* K-12	700 coding fragments and 700 convergent intergenic fragments of *E. coli* K-12
Sequence length	81bp (-60 to +20 bp flanking TSS. TSSs are at the 0th site.)	81bp
Sample size	741	1400

Meanwhile, the negative samples including 1400 none-promoter were randomly extracted from coding regions and intergenic regions of *E. coli* K-12 genome. The length of all negative sequences is also 81 bp, which is the same as that of the positive samples. Details of positive and negative samples are shown in Table 1.

Finally, 741 positive samples and 1400 negative samples were obtained for the benchmark dataset *S*, which can be expressed by

$$S = S^{+} \bigcup S^{-} \tag{1}$$

where S^{+} represents positive samples or promoter sequences, S^{-} represents negative samples or non-promoter sequences, and the symbol ∪ represents the "union" in the set theory.

2.2 Formulation of DNA Samples

A DNA segment with L nucleic acid residues can be formulated as,

$$\boldsymbol{D} = R_1R_2R_3R_4R_5R_6R_7 \ldots R_L \tag{2}$$

where $R_i\,(i = 1, 2, 3, \ldots, L)$ represents the nucleic acid residue at position *i* in the DNA sequence sample ***D***, and *L* represents the length of this DNA sequence. Since the input format of data has special requirement for transforming the DNA segment as input data type for statistical prediction, the primal problem is how to extract features effectively, and express the DNA segment correctly with a discrete model or a vector.

The existing method [3], [9], [10] only considers single base and oligodeoxynucleotide, which would be disable to transform longer motifs as features for predicting promoter. As a bioinformatics arithmetic for genome analysis, the Z-curve method, proposed initially by Zhang et al. [11], [12], can obtain sequence features from nucleotide sequence through calculating the frequencies of single DNA base (A, C, G, T) and mapping the sequence to three-dimensional space (*x*-axis, *y*-axis and *z*-axis). Z-curve can express the whole DNA sequence in a mathematical way. Therefore, we introduced a novel modified Z-curve named "multi-window Z-curve", which can represent the frequency features and three dimensionality features of diverse length sequences. For the sake of practical purposes and reducing the computation complexity, we set $w = 1, 2, \ldots, 7$ (ω is the window size). The details of formulas are expressed as follows:

$$\begin{cases} x_i = \left(P_{S_{\omega-1}A} + P_{S_{\omega-1}G}\right) - \left(P_{S_{\omega-1}C} + P_{S_{\omega-1}T}\right) \\ y_i = \left(P_{S_{\omega-1}A} + P_{S_{\omega-1}C}\right) - \left(P_{S_{\omega-1}G} + P_{S_{\omega-1}T}\right) \\ z_i = \left(P_{S_{\omega-1}A} + P_{S_{\omega-1}T}\right) - \left(P_{S_{\omega-1}C} + P_{S_{\omega-1}G}\right) \end{cases} \\ \left(w = 1, 2, \ldots, 7; i = 1, 2, \ldots, 4^{w-1}\right) \tag{3}$$

where $i = 1, 2, \ldots, 4^{\omega-1}$, w is the window size, $S_{\omega-1}$ is the string of selected bases, $P_{S_{\omega-1}x}$ (*x* = *A*, *T*, *C*, *G*) is the corresponding frequency of string $S_{\omega-1}x$. For example, when $\omega = 2$, S_1= *A*, *C*, *G*, *T*, and P_{S_1x} are the corresponding frequencies of P_{AA},P_{AC}, P_{AG}, P_{AT}, ..., P_{TT}, respectively. Thus, the bases with the length from 1 to 7 are all taken into account in the Eq. (3), and when $\omega = 0$, the Eq. (3) is just the original form of Z-curve.

In addition, the global-range or the long-range correlation information was also used as the features in the nucleosome positioning prediction [13] and human origin of replication prediction [14]. Stimulated by the successful application of the 'pseudo K-tuple nucleotide composition' or PseKNC [15] in identifying $\boldsymbol{\sigma}^{54}$ promoter, the correlation factor of PseKNC was also used to describe $\boldsymbol{\sigma}^{70}$ promoter in the current work as follows.

The *j*-th tire correlation factor θ_j displays the sequence order correlation between all of the *j*-th most contiguous dinucleotides along a DNA sequence and can be formulated by

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta\left(R_iR_{i+1}; R_{i+j}R_{i+j+1}\right) \\ (j = 1, 2, \cdots, \lambda, \text{here } \lambda < L) \tag{4}$$

where λ is the number of the total counted ranks(or tiers) of the long-range correlations along a DNA sequence. The correlation function $\Theta(R_iR_{i+1}; R_{i+j}R_{i+j+1})$ in Eq. (4) is defined by

$$\Theta\left(R_iR_{i+1}; R_{i+j}R_{i+j+1}\right) \\ = \frac{1}{\mu}\sum_{v=1}^{\mu}\left[P_v(R_iR_{i+1}) - P_v\left(R_{i+j}R_{i+j+1}\right)\right]^2 \tag{5}$$

where μ is the number of local DNA structural properties and it equals to 6 in the current research that will be explained in the Section 2.3; $P_v(R_iR_{i+1})$ is the numerical value of the ***v***-th ($\nu = 1, 2, \cdots, \mu$) DNA local structural property for the dinucleotide R_iR_{i+1} at position *i* in DNA sequence and $P_v(R_{i+j}R_{i+j+1})$ is the value for the dinucleotide $R_{i+j}R_{i+j+1}$ at position *i*+*j* in the same sequence. It will be given in Eq. (9).

By combining multi-window Z-curve with PseKNC, a new formulation, called 'pseudo multi-window Z-curve nucleotide composition' or 'PseZNC', is given by

$$\mathbf{D}_{\text{PseZNC}} = \left[d_1\ d_2 \ldots d_{3\times4^{\omega-1}}\ d_{3\times4^{\omega-1}+1} \ldots d_{3\times4^{\omega-1}+\lambda}\right]^T \tag{6}$$

in which

$$d_u = \begin{cases} f_{x_1}^{1}\ f_{y_1}^{2}\ f_{z_1}^{3} \ldots f_{x_{4^{\omega-1}}}^{3\times4^{\omega-1}-2}\ f_{y_{4^{\omega-1}}}^{3\times4^{\omega-1}-1}\ f_{z_{4^{\omega-1}}}^{3\times4^{\omega-1}} & (1 \le u \le 3\times4^{\omega-1}) \\ \theta_1\ \theta_2\ \theta_3 \ldots\ \theta_\lambda & (3\times4^{\omega-1}+1 \le u \le 3\times4^{\omega-1}+\lambda) \end{cases} \tag{7}$$

where $d_u(1 \le u \le 3\times4^{\omega-1})$ are the features extracted from the method of 'multi-window Z-curve', and each element *f* is the feature value which represents the frequency of bases, and it can be mapped into x, y, x-axis using the Eq.(3). $d_u(3 \times 4^{\omega-1}+ 1 \leq u \leq 3 \times 4^{\omega-1} + \lambda)$ are the features extracted from the correlation factor of PseKNC, and each element θ can be calculated by using the Eqs.(4-5).

2.3 Parameters of DNA Local Structural Property

Many researches have revealed the fact that DNA local structural properties are of great significance in many biological processes, such as protein-DNA interactions [16], nucleosome occupancy [17], the formation of chromosomes [18], research on meiotic recombination [19], and so on. Promoters, acting as important and particular regulators, have some characteristic DNA structural properties to recruit special binding proteins and trigger further gene expression and regulation. In addition, many articles have built models to capture a great deal of the complex structural features of chromatin. These works displayed that physicochemical properties did play an important role in the recognition of promoters [6], [16], [18]. Moreover, Duran et al. [20] have strongly held the view that the ancient regulatory mechanism determined by the intrinsic physical properties of the DNA may lead to the complicacy of transcription regulation in human genome.

In general, the spatial contribution of two consecutive base pairs can be tokened by six parameters, in which three of them are local translational characterizations and the remaining three are the local angular ones. They were formulated by the following equation:

$$\text{Translational} = \begin{cases} \text{Slide} \\ \text{Shift} \\ \text{Rise} \end{cases} \text{Angular} = \begin{cases} \text{Roll} \\ \text{Tilt} \\ \text{Twist} \end{cases} \tag{8}$$

These values which can be found from [13] will be used to compute the global or long-range sequence-order information of the promoter sequences through Eqs. (4), (5).

However, before we substituted the values of physicochemical properties into Eqs. (4), (5), all the original values for $P_v(R_iR_{i+1})(v = 1, 2, \cdots, 6)$ were contingent on a standard conversion as expressed by the following equation:

$$P_v(R_iR_{i+1}) \Leftarrow \frac{P_v(R_iR_{i+1}) - \langle P_v(R_iR_{i+1})\rangle}{\text{SD}\langle P_v(R_iR_{i+1})\rangle} \tag{9}$$

where the symbol <> represents the average of the quantity over the 16 different combinations of *A*, *C*, *G*, *T* for R_iR_{i+1}, and SD is the corresponding standard deviation [19]. The derived values obtained by Eq. (9) will have a mean value of zero over the 16 diverse dinucleotides, and will sustain unchanged if the same conversion procedure is implemented again.

2.4 Support Vector Machine

Support vector machine (SVM) is a useful and effective supervised machine learning algorithm for sample classification and pattern recognition, which has been successfully applied in the field of bioinformatics [21], [22], [23]. SVM can map the input vectors into high-dimensional feature spaces through kernel function and construct a hyperplane or set of hyperplanes in the high dimensional space. More detailed descriptions about the formulation of SVM and how it works can be seen in the reference articles [22], [24], [25]. In this study, the LIBSVM 3.20 package [26] was used to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis function (RBF) was selected as the kernel function in this study. In order to obtain the optimal model, a grid search method was implemented to gain the regularization parameter C and the kernel width parameter γ.

2.5 Performance Evaluation

The 5-fold cross-validation was used for evaluating the performance of the predictor. The following four metrics were used to measure the proposed predictor.

$$S_n(\text{Sensitivity}) = \frac{TP}{TP + FN} \tag{10}$$

$$S_p(\text{Specificity}) = \frac{TN}{TN + FP} \tag{11}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{12}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

In the Eqs (10-13), *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives and *FN* is the number of false negatives. In addition, S_n indicates the capability of correctly identifying positive samples, S_p indicates the capability of correctly predicting negative samples, *Acc* and *MCC* stand for the accuracy and Mathew's correlation coefficient, respectively.

2.6 Feature Selection

Along with the increase of ω and λ, the dimension of $\mathbf{D}_{\text{PseZNC}}$ in Eq. (6) will increase dramatically. Generally, a large number of features will result in the high-dimension disaster in the following three negative aspects: (i) overfitting that will cause a serious bias and extremely low generalization ability for the predictor; (ii) noise or the information redundancy that will lead to very poor prediction accuracy; (iii) time-consuming which will occupy a lot of computer resources.

In this study, we performed feature selection using the *F*-score algorithm. The *F*-score of *i*-th feature is defined by:

$$F_i = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^{(+)}}\left(\bar{x}_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n^--1}\sum_{k=1}^{n^{(-)}}\left(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \tag{14}$$

where n^+ and n^- are the total number of the positive samples and the negative samples, respectively; $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the mean value of the *i-th* feature of the entire positive samples and the entire negative samples, respectively; while $\bar{x}_i$ represents the mean value of the total samples; $\bar{x}_{k,i}{}^{(+)}$ and $\bar{x}_{k,i}{}^{(-)}$ represent the *i-th* feature of the *k-th* sample in the positive dataset and negative dataset, respectively.

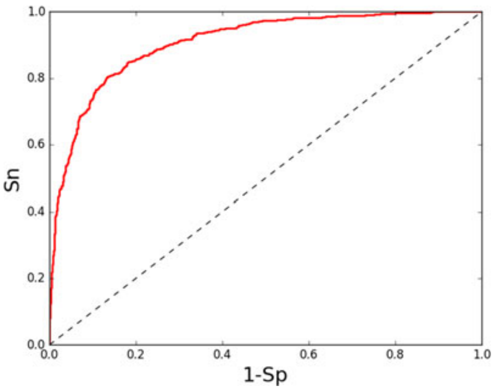


Fig. 1. The ROC curve for the model with 894 optimal features in the 5-fold cross-validation. The diagonal dotted line denotes a random guess. The AUC is 0.9088 for our model.

As mentioned above, the F-scores of all features can be obtained by this feature selection method. The higher the F-score of feature is, a better discriminative capability it will has. Thus, all features were ranked in their descending order based on their F-score values. The Incremental Feature Selection (IFS) was used to determine the optimal number of features [27]. Then we acquired a series of feature subsets through appending a feature one by one in line with the descending order. Eventually, the N feature-subsets were constructed and formulated as

$$S_\varepsilon = \{F_1\ F_2 \cdots F_\varepsilon\}(1 \le \varepsilon \le N) \tag{15}$$

where N is the number of total features, F_ε is the feature with the ε-th highest F-score value. For each of such feature-subsets, we constructed a SVM prediction model assessed by the 5-fold cross-validation test on the benchmark dataset.

3 Results and Discussions

3.1 Parameter Optimization

Through the Eqs (6), (7), the prediction results depended on two parameters, ω and λ, where ω is the window size of 'multi-window Z-curve', that reflects the effect of the local or short-range sequence order; and λ is the number of the total counted ranks (or tiers) of the physicochemical properties correlations along a DNA sequence, that reflects the effect of the global or long-range sequence order. In general, the larger the ω and λ are, the greater number of features of the local-range and global-range sequence the model contains. However, if ω or λ is too large, it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy due to overfitting or 'high dimension disaster' problem as mentioned above. Hence, to obtain the optimal values of the two parameters, we set a search strategy as follows:

$$\begin{cases} 2 \le \omega \le 7 & \text{with step } \Delta = 1 \\ 1 \le \lambda \le 70 & \text{with step } \Delta = 1 \end{cases} \tag{16}$$

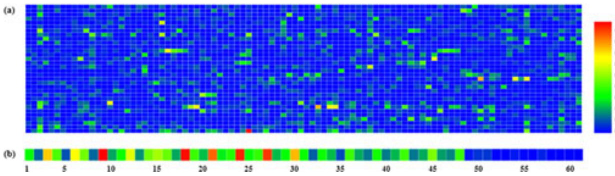


Fig. 2. A heat map for displaying the features according to their F-score value. (a) A heat map was used for representing the 3,072 features that were extracted by multi-window Z-curve, which reflect the short-range or local-range sequence information. (b) A heat map was used for representing the 61 features extracted by PseKNC, which reflect the long-range or global-range sequence information.

As for the search strategy, a total of 6×70 = 420 combinations should be taken into account when we optimized the two parameters. Finally, we obtained the prediction results which summarized as

$$\begin{cases} S_n = 74.63\% \\ S_p = 79.50\% \\ Acc = 77.81\% \\ MCC = 0.5274 \end{cases} \quad \text{when}\begin{pmatrix} \omega = 6 \\ \lambda = 61 \end{pmatrix} \tag{17}$$

3.2 Feature Optimization

According to Eqs (3), (4), (5), (6), and (7), when $\omega = 6$ and $\lambda = 61$, there are $3 \times 4^5 + 61 = 3133$ features. The F-score algorithm was used to further improve prediction accuracy.

By virtue of the feature selection procedure mentioned above, we found that the best prediction results can be obtained by using 894 optimal features and were shown as follows:

$$\begin{cases} S_n = 80.30\% \\ S_p = 86.79\% \\ Acc = 84.54\% \\ MCC = 0.6631 \end{cases} \quad (894\ optimal\ features) \tag{18}$$

In this prediction model, the key components for the feature vector has $841 + 53 = 894$ features, in which 841 features reflect the effect of the local or short range sequence order information and the other 53 features represent the effect of the global or long range sequence order information.

Meanwhile, to provide a graphical illustration to show the performance of the current binary classifier, a 2D plot, called ROC (receiver operating characteristic) curve was ploted in Fig. 1. The ROC curve can measure the predictive capability of our method across the entire range of SVM decision values. The area under the ROC (AUC) was calculated to quantitatively and objectively measure the performance of the proposed method. A perfect classifier gives AUC = 1, and the random performance gives AUC = 0.5. The AUC of our model is 0.9088 indicating that the proposed prediction model is quite powerful.

3.3 Feature Analysis

In order to analyze the contributions of different components in this predictor, a heat map was drawn in Fig. 2 for the graphical representation of each feature with different color according to its F-score value that has been scaled between 0 and 1. From Fig. 2(a), the majority of all 3072 features are blue, which means that most of these features are irrelevant with the promoter recognition. By analyzing the highly relevant hexamers (features lean towards red), we

TABLE 2
Comparative Results of Two Methods for Identifying $\boldsymbol{\sigma}^{70}$ Promoters

	S_n	S_p	*Acc*	*MCC*	AUC
Z-curve	74.6%	79.5%	77.8%	0.527	0.848
PseZNC	**80.3%**	**86.8%**	**84.5%**	**0.663**	**0.909**
IPMD	82.4%	90.7%	87.9%	0.731	—

obtained some important hexamers which play key roles in recognizing promoters such as 'TTGTAX', 'GCCGGX' and 'TATAAX' (X is A, C, G or T). Further analysis demonstrated that all of these three key features for identifying promoters are derived from the dimensionality information of multi-window Z-curve, and the value of feature f_{3015}, f_{1807} and f_{2451} are listed as follows:

$$\begin{cases} f_{3015} = (f_{\text{TTGTAA}} + f_{\text{TTGTAT}}) - (f_{TTGTAC} + f_{TTGTAG}) \\ f_{1807} = (f_{GCCGGA} + f_{GCCGGG}) - (f_{GCCGGC} + f_{GCCGGT}) \\ f_{2451} = (f_{TATAAA} + f_{TATAAT}) - (f_{TATAAC} + f_{TATAAG}) \end{cases} \tag{19}$$

where f_{3015}, f_{1807} and f_{2451} are calculated through the simple addition and subtraction by the multi-window Z-curve in Eq. (3).

In the Fig. 2 (b), it shows the heatmap of 61 features of PseKNC that represents the long-range or global-range sequence information. Then we found the top 5 features which leap to red are 9, 18, 24, 27 and 21. The five factors have the highest values of *F*-score, and are superior to others for identifying the $\boldsymbol{\sigma}^{70}$ promoter.

3.4 Comparison with Method without Pseudo Nucleotide Composition

Many previous methods [3], [9], [10] only extracted short-range or local-range sequence information as features to recognize prokaryotic promoter. While in this study, we considered both short-range sequence information extracted through multi-window Z-curve and long-range sequence information extracted through PseKNC to predict σ^{70} promoters. In order to demonstrate the importance of long-range sequence information, we investigated the performance of features calculated by multi-window Z-curve features. The results were listed in Table 2. It demonstrates that the long-range information also plays important role in $\boldsymbol{\sigma}^{70}$ promoter prediction.

Song [28] used the multi-window Z-curve to predict $\boldsymbol{\sigma}^{70}$ promoters and reported a very high accuracy. She compared Z-curve-based method with the IPMD-based method [3], and claimed that her method was better than IPMD. However, we think the comparsion was cursory, arbitrary and not objective because of the use of different benchmark datasets. To make a objective comparison, we extracted multi-window Z-curve and PseZNC from same benchmark dataset. Results in Table 2 show that the PseZNC-based method is superior to multi-window Z-curve-based method. The IPMD-based method achieved the highest accuracy among three methods. However, it did not take long-range sequence information into account, which would miss some important characteristics for the correct identification of $\boldsymbol{\sigma}^{70}$ promoters. The current method may play a complementary role to other existing methods for predicting $\boldsymbol{\sigma}^{70}$ promoters.

Using the multi-window Z-curve and pseudo oligonucleotide composition to incorporate, respectively, the local and global sequence-order informations, a predictor called iPro70-PseZNC which can be freely accessible at http://lin.uestc.edu.cn/server/iPro70-PseZNC was developed for identifying the $\boldsymbol{\sigma}^{70}$ promoters. In the predictor, the feature selection technique was used to winnow out the key features. It was observed that the key features thus obtained did really represent the regulatory motifs in $\boldsymbol{\sigma}^{70}$ promoter sequences. The PseZNC method proposed in this study can also be generalized to the identification of other DNA regulatory elements.

Acknowledgments

This work was supported by the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and 14JC0121), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015J144; ZYGX2015Z006), Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), China Postdoctoral Science Foundation (No.2015M582533), and the Scientific Research Foundation of the Education Department of the Sichuan Province (11ZB122).

References

[1] P. Smolen, D. A. Baxter, and J. H. Byrne, "Modeling transcriptional control in gene networks–methods, recent results, and future directions," *Bull. Math. Biol.*, vol. 62, pp. 247–292, Mar. 2000.

[2] J. Hasty, F. Isaacs, M. Dolnik, D. McMillen, and J. J. Collins, "Designer gene networks: Towards fundamental cellular control," *Chaos*, vol. 11, pp. 207–220, Mar. 2001.

[3] H. Lin and Q. Z. Li, "Eukaryotic and prokaryotic promoter prediction using hybrid approach," *Theory Biosci.*, vol. 130, pp. 91–100, Jun. 2011.

[4] B. Demeler and G. W. Zhou, "Neural network optimization for E. coli promoter prediction," *Nucleic Acids Res.*, vol. 19, pp. 1593–1599, Apr. 11, 1991.

[5] R. R. Mallios, D. M. Ojcius, and D. H. Ardell, "An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis sigma66 promoters," *BMC Bioinf.*, vol. 10, 2009, Art. no. 271.

[6] Y. C. Zuo and Q. Z. Li, "The hidden physical codes for modulating the prokaryotic transcription initiation," *Physica A: Stat. Mechanics Appl.*, vol. 389, pp. 4217–4223, 2010.

[7] Q. Wu, J. Wang, and H. Yan, "An improved position weight matrix method based on an entropy measure for the recognition of prokaryotic promoters," *Int. J. Data Min. Bioinf.*, vol. 5, pp. 22–37, 2011.

[8] S. Gama-Castro, et al., "RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Res.*, vol. 44, pp. D133–D143, Jan. 4, 2016.

[9] P. B. Horton and M. Kanehisa, "An assessment of neural network and statistical approaches for prediction of E. coli promoter sites," *Nucleic Acids Res.*, vol. 20, pp. 4331–4338, Aug. 25, 1992.

[10] A. M. Huerta and J. Collado-Vides, "Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals," *J. Mol. Biol.*, vol. 333, pp. 261–278, Oct 17, 2003.

[11] C. T. Zhang, "A symmetrical theory of DNA sequences and its applications," *J. Theor Biol.*, vol. 187, pp. 297–306, Aug 7, 1997.

[12] C. T. Zhang, R. Zhang, and H. Y. Ou, "The Z curve database: a graphic representation of genome sequences," *Bioinf.*, vol. 19, pp. 593–599, Mar. 22, 2003.

[13] S. H. Guo, et al., "iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinf.*, vol. 30, pp. 1522–1529, Jun. 1, 2014.

[14] C. J. Zhang, H. Tang, W. C. Li, H. Lin, W. Chen, and K. C. Chou, "iOri-Human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition," *Oncotarget*, vol. 7, pp. 69783–69793, Sep. 12, 2016.

[15] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K. C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Res.*, vol. 42, pp. 12961–12972, Dec. 1, 2014.

[16] J. R. Goni, A. Perez, D. Torrents, and M. Orozco, "Determining promoter location based on DNA structure first-principles calculations," *Genome Biol.*, vol. 8, 2007, Art. no. R263.

[17] V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes, and T. Grange, "DNA physical properties determine nucleosome occupancy from yeast to fly," *Nucleic Acids Res.*, vol. 36, pp. 3746–3756, Jun. 2008.

[18] J. R. Goni, C. Fenollosa, A. Perez, D. Torrents, and M. Orozco, "DNAlive: A tool for the physical analysis of DNA at the genomic scale," *Bioinf.*, vol. 24, pp. 1731–1732, Aug. 1, 2008.

[19] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Anal. Biochem.*, vol. 370, pp. 1–16, Nov. 1, 2007.

[20] E. Duran, et al., "Unravelling the hidden DNA structural/physical code provides novel insights on promoter location," *Nucleic Acids Res.*, vol. 41, pp. 7220–7230, Aug. 2013.

[21] S. Q. Wang, J. Yang, and K. C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition," *J. Theoretical Biol.*, vol. 242, pp. 941–946, 2006.

[22] K. C. Chou and Y. D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *J. Biol. Chemistry*, vol. 277, pp. 45765–45769, 2002.

[23] P. M. Feng, W. Chen, H. Lin, and K. C. Chou, "iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Anal. Biochem.*, vol. 442, pp. 118–125, Nov. 1, 2013.

[24] Y. D. Cai, G. P. Zhou, and K. C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical J.*, vol. 84, pp. 3257–3263, 2003.

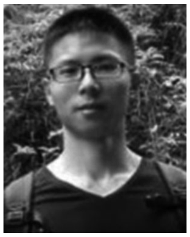
[25] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinf.*, vol. 16, pp. 906–914, Oct. 2000.

[26] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, 2011, Art. no. 27.

[27] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Mol. Biosyst.*, vol. 12, pp. 1269–1275, Apr 22, 2016.

[28] K. Song, "Recognition of prokaryotic promoters based on a novel variable-window Z-curve method," *Nucleic Acids Res.*, vol. 40, pp. 963–971, Feb. 2012.

Hao Lin received the PhD degree in biophysics from Inner Mongolia University, in 2007. He is a professor of the Center for Informational Biology at the University of Electronic Science and Technology of China. His research is in the areas of bioinformatics and systems biology.



Zhi-Yong Liang is working toward the master's degree in the Center for Informational Biology at the University of Electronic Science and Technology of China. His research interests include bioinformatics.

Hua Tang is an associate professor of the Department of Pathophysiology at Southwest Medical University. Her research is in the areas of bioinformatics and pathophysiology.

Wei Chen received the PhD degree from Inner Mongolia University in 2010. He is a professor of the School of Science at the North China University of Science and Technology. His research is in the areas of bioinformatics.