Molecular Therapy
Nucleic Acids
Original Article

iProEP: A Computational Predictor for Predicting Promoter

Hong-Yan Lai,[1,4] Zhao-Yue Zhang,[1,4] Zhen-Dong Su,[1] Wei Su,[1] Hui Ding,[1] Wei Chen,[1,2,3] and Hao Lin[1]

[1]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; [2]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China; [3]Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China

Promoter is a fundamental DNA element located around the transcription start site (TSS) and could regulate gene transcription. Promoter recognition is of great significance in determining transcription units, studying gene structure, analyzing gene regulation mechanisms, and annotating gene functional information. Many models have already been proposed to predict promoters. However, the performances of these methods still need to be improved. In this work, we combined pseudo *k*-tuple nucleotide composition (PseKNC) with position-correlation scoring function (PCSF) to formulate promoter sequences of *Homo sapiens* (*H. sapiens*), *Drosophila melanogaster* (*D. melanogaster*), *Caenorhabditis elegans* (*C. elegans*), *Bacillus subtilis* (*B. subtilis*), and *Escherichia coli* (*E. coli*). Minimum Redundancy Maximum Relevance (mRMR) algorithm and increment feature selection strategy were then adopted to find out optimal feature subsets. Support vector machine (SVM) was used to distinguish between promoters and non-promoters. In the 10-fold cross-validation test, accuracies of 93.3%, 93.9%, 95.7%, 95.2%, and 93.1% were obtained for *H. sapiens*, *D. melanogaster*, *C. elegans*, *B. subtilis*, and *E. coli*, with the areas under receiver operating curves (AUCs) of 0.974, 0.975, 0.981, 0.988, and 0.976, respectively. Comparative results demonstrated that our method outperforms existing methods for identifying promoters. An online web server was established that can be freely accessed (http://lin-group.cn/server/iProEP/).

INTRODUCTION

In a genome, promoters are important regions of DNA that locate near the transcription start sites (TSSs) of genes.[1] They are essentially nucleotide sequences of approximately extending dozens to hundreds base pairs upstream and downstream of the TSS. They always serve as regulatory elements for the assembly of transcription machinery, especially combining with RNA polymerase[2] for promoting accurate initiation of transcription. Additionally, evidence has proved that promoters play crucial roles in the regulation of gene expression, such as alternative splicing, stability of transcripts, mRNA localization, and translation.[3] The identification of promoters in a gene is an important part of the recognition of a gene's complete structure. Hence, the mapping of promoters to genome is usually the first step in unraveling the mechanisms of gene transcriptional and expressional regulation. Therefore, research on promoter prediction is full of significance and deserves to be pushed forward.

DNA elements in promoters are different between eukaryotes and prokaryotes. In eukaryotes, most protein-coding genes and some nuclear small RNAs have binding sites for RNA polymerase II. The core region of RNA polymerase II-dependent promoters usually contains several regulatory units: the TATA element, which is located 25 bp upstream of the TSS; the initiator; and the downstream promoter element (DPE), usually located 30 bp downstream of the TSS.[4] In prokaryotes, most genes are regulated by the σ^{70} promoter, which contains three basic elements: the Pribnow Box with the consensus sequence 5′-TATAAT-3′ located 10 bp upstream of the TSS, the −35 region with the consensus sequence 5′-TTGACA-3′ located 35 bp upstream of the TSS, and the initiator adjacent to the TSS.[5,6] Distinct gene-regulatory mechanisms and sequence compositions among species promote us to use different methods to identify promoters in their genomes.[7,8]

With the development of high-throughput sequencing technology, increasing genomes need to be annotated. It is costly, laborious, and time consuming to use experimental methods to characterize promoters, however, which promotes the development of the computational methods in promoter identification. There have been many attempts to predict promoters in different species. Some models were based on the principle of sequence similarity, and others

Received 23 February 2019; accepted 19 May 2019;
https://doi.org/10.1016/j.omtn.2019.05.028.

[4]Senior author

Correspondence: Wei Chen, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China.
E-mail: chenweiimu@gmail.com
Correspondence: Hao Lin, 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

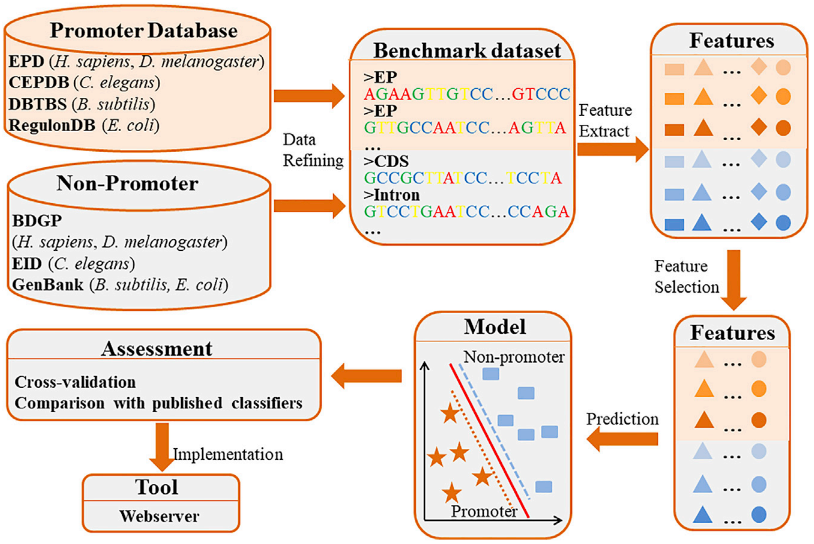


Figure 1. A Flowchart to Outline the Promoter Prediction Program Construction

converted the original sequences into numeric sequences and then adopted machine learning approaches to perform recognition. The latter extracted features according to various promoter properties, such as CpG content,[9] free energy,[10] consensus sequence,[11] and global descriptor,[10] and built the prediction programs based on machine learning approaches, such as Fisher's linear discriminant,[10] decision tree,[12] support vector machine (SVM),[13] Hidden Markov Model,[11] neural network,[14] pattern-based nearest neighbor search approach,[15] and so on. Recently, deep learning has been used to grasp complex promoter sequence characteristics[16,17] and related bioinformatics identification problems.[18–22] Although existing algorithms have exhibited encouraging performance, most of those predictors focused on only one species, and there is still space for prediction performance improvement.

In this study, according to the steps shown in Figure 1, we developed an effective and powerful computational promoter prediction program for eukaryote and prokaryote species. We firstly collected promoter and non-promoter sequences in five species to construct the reliable benchmark datasets. The features extracted from the primary sequences were filtered according to the ability of distinguishing promoters from non-promoters by using feature selection technique. Subsequently, the optimal features were inputed into the SVM to train, test, and build models. Finally, based on the proposed model, we established a user-friendly web server iProEP, which can be freely accessed at http://lin-group.cn/server/iProEP/.

Table 1. The Optimal Values of Three PseKNC Parameters for Five Species

Kingdom	Species	k	λ	ω	ACC (%)
Eukaryotes	*H. sapiens*	4	24	0.1	90.9
	D. melanogaster	5	9	0.1	89.5
	C. elegans	4	22	0.1	81.4
Prokaryotes	*B. subtilis*	4	12	0.2	83.8
	E. coli	4	12	0.1	80.7

RESULTS

Optimization of Three PseKNC-Related Parameters

As indicated in the PseKNC section (see Materials and Methods), three parameters, k, λ, and ω, must be determined when using PseKNC to formulate promoters and non-promoters. In the PseKNC, the k and λ describe the short-range and long-range sequence-order effect, respectively, and ω is the weight factor to adjust the ratio of the two effects. In this work, the optimal values of the three parameters for five species can be obtained by searching the following scopes:

$$\begin{cases} k \in [2,\ 6], & step = 1 \\ \lambda \in [1,\ 30], & step = 1 \\ \omega \in [0.1,\ 1], & step = 0.1 \end{cases} . \qquad \text{(Equation 1)}$$

For each species, the performances of 1,500 (5 × 30 × 10) different combinations of three parameters were examined to obtain their optimal combination that could produce best accuracy. Thus, we constructed 1,500 SVM classifiers based on 5-fold cross-validation for each species. The optimal combinations of the three parameters for five species were reported in Table 1.

The Ultimate Five Promoter Classifiers

By combining PseKNC with position-correlation scoring function (PCSF), promoter and non-promoter samples can be formulated by $(4^k + 6\lambda + n)$ dimension features. In the $4^k + 6\lambda$ dimension PseKNC features, 4^k reflects the DNA short-range correlation information, and 6λ describes the long-range correlation information. The position information is characterized by n dimension PCSF (see Materials and Methods). When incorporating these features into a prediction model, redundant information or noise might influence the performance of the model. Therefore, Minimum Redundancy Maximum Relevance (mRMR) combined with the increment feature selection (IFS) process was adopted to eliminate these unrelated features for improving the accuracy and robustness of promoter recognition models.

Ultimately, by constructing a great number of SVM-based models and comparing these models' performance using 5-fold

Table 2. The Feature Numbers and Accuracies for Five Species before and after mRMR Feature Selection

		Original Features		Optimal Features	
Kingdom	Species	Feature Number	*ACC* (%)	Feature Number	*ACC* (%)
Eukaryotes	*H. sapiens*	423	93.4	410	93.5
	D. melanogaster	1,097	93.3	893	93.8
	C. elegans	405	94.4	65	95.6
Prokaryotes	*B. subtilis*	345	94.0	55	95.5
	E. coli	345	92.1	44	93.2

Table 3. The Results for Five Species by Using 10-Fold Cross-Validation

Kingdom	Species	*ACC* (%)	*Sn* (%)	*Sp* (%)	*AUC*
Eukaryotes	*H. sapiens*	93.3	92.3	92.7	0.974
	D. melanogaster	93.9	92.6	92.6	0.975
	C. elegans	95.7	95.0	94.4	0.981
Prokaryotes	*B. subtilis*	95.2	94.8	94.3	0.988
	E. coli	93.1	92.2	91.2	0.976

cross-validation, the optimal feature subsets for five species were screened out and shown in Table 2. It is obvious that the accuracies were indeed improved after removing noise features. It was also noted that the feature dimensions for *C. elegans*, *B. subtilis*, and *E. coli* were dramatically decreased after feature selection. However, only 13 and 204 features were excluded for *H. sapiens* and *D. melanogaster*. The reason for these phenomena may be that promoter sequences of *H. sapiens* and *D. melanogaster* are much more complex than those of the other three species.

After determining the optimal feature subsets, for convenience in subsequent comparisons, the 10-fold cross-validation was applied to seek the best SVM-related parameters (c and γ) and to evaluate those models. For *H. sapiens*, *D. melanogaster*, *C. elegans*, *B. subtilis*, and *E. coli*, the optimal values of c and γ are 2 and 2^{-3}, 2 and 2^{-2}, 2^5 and 2^{-1}, 2^5 and 2^{-7}, and 2^{-1} and 2^{-1}, respectively. The detailed results were listed in Table 3. In addition, receiver operating characteristics (ROC) curves were also plotted in Figure 2 to visually show the prediction capability of our model on discrimination between promoters and non-promoters.

Comparison with Existing Promoter Classifiers

Comparison with other existing methods is an important strategy to highlight the merits of proposed models. Currently, several computational methods have been developed for eukaryote and prokaryote promoter prediction.[17,23] To provide a fair comparison of the same data, only a method called IPMD[24] was used to make comparisons, because the same benchmark datasets and same cross-validation rule were used in both works. Furthermore, comparison in the paper[24] has demonstrated that IPMD is superior to other existing predictors, such as NNPP2.2, McPromoter. The IPMD is a hybrid method that combined PCSF and increment of diversity (ID) with the modified Mahalanobis Discriminant. Figure 3 recorded the results obtained by our proposed method and IPMD. The results show that our model is superior to the IPMD model, especially for *C. elegans*, *B. subtilis*, and *E. coli*.

Moreover, multi-window Z-curve[25] and PseZNC[26] have been proposed as feature extraction approaches for σ^{70} promoter prediction in *E. coli*. Based on the same *E. coli* data, multi-window Z-curve was re-evaluated in Lin et al.[26] Its overall accuracy is only 77.81% with the area under receiver operating curve (AUC) of 0.8480, which is lower than those of our proposed method. PseZNC is a feature extraction technique that combines multi-window Z-curve with PseKNC. The accuracy of the PseZNC-based method is also lower than our method. Detailed comparison was exhibited in Figure 4. Z-curve theory has been successfully applied in prokaryotic gene prediction because of the characteristics of period-3 in codon. However, promoter sequence cannot code amino acids and dose not obey the codon rule. This is why the two Z-curve-based methods cannot produce better results on promoter prediction.

Recently, two predictors called iPromoter-2L[27] and MULTiPly[28] were also designed for *E. coli* promoter prediction. We could make a raw comparison because the benchmark data in these studies were all derived from RegulonDB. Both predictors could provide multi-layer prediction for recognizing promoters and their subtypes. The former was based on multi-window-based PseKNC and Random Forest, which produced the accuracy (*ACC*), sensitivity (*Sn*), and specificity (*Sp*) of 81.68%, 79.20%, and 84.16%, respectively. The latter obtained the related three indexes of 86.92%, 87.27%, and 86.57% by a SVM-based model. It was found that our proposed model yielded *ACC*, *Sn*, and *Sp* of 93.1%, 92.2%, and 91.2%, respectively (Table 3), which are superior to the two predictors.

Cross-Species Evaluation

Cross-species evaluation on eukaryote and prokaryote was performed to assess the generalization ability of the proposed method. It should be noted that because of the different sequence structure, composition, and regulatory mechanism between eukaryote and prokaryote, the following experiments were performed. We first evaluated the *H. sapiens*-based model on *D. melanogaster* and *C. elegans* data. Results (Table 4) showed that the accuracies are only 77.10% and 66.63% for the two test datasets. Subsequently, we investigated the prediction performances of the *D. melanogaster*-based model on *H. sapiens* and *C. elegans* data. Only 68.41% of *H. sapiens* sequences and 65.68% of *C. elegans* sequences can be correctly identified. Finally, we performed similar examinations and obtained similar results on the models from *C. elegans*, *B. subtilis*, and *E. coli*. The unsatisfactory results are mainly due to the species-specificity property of promoter sequences.

Web Server and Tutorial

A user-friendly and publicly accessible web server could provide convenience for researchers.[29–31] Thus, based on our proposed method,

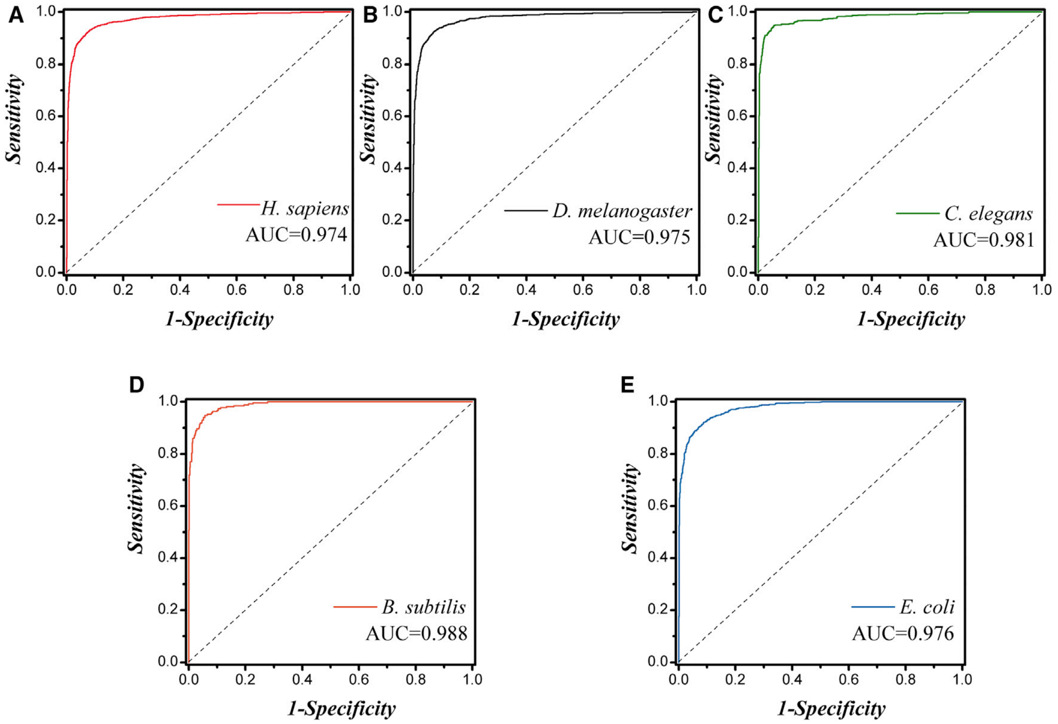


Figure 2. Evaluating the iProEP by Using ROC Curve

ROC curves for promoter prediction in (A) *H. sapiens*, (B) *D. melanogaster*, (C) *C. elegans*, (D) *B. subtilis*, and (E) *E. coli*.

we established a powerful web server called iProEP, by which researchers can identify promoters by uploading DNA sequences. A step-by-step guide on how to use the web server is given as follows:

Step 1. Click on the web address http://lin-group.cn/server/iProEP/ and the user will see the brief summary about iProEP (Figure 5).

Step 2. Click on the "Predictor" on the navigation bar, then choose a suitable species and input the query DNA sequences into the input box for prediction. It should be noted that the sequences must be FASTA format with the length of >300 bp for eukaryote and >81 bp for prokaryote. Click on the "example" button below the input box to see the sample sequence in the FASTA format.

Step 3. Click on the "submit" button to obtain the predicted result. If the sequence is longer than 300 or 81 bp, the predictor will scan the sequence using the 300- or 81-bp window with the step of 1 bp for eukaryote or prokaryote, respectively. The result for each subsequence will be displayed on the result page.

DISCUSSION

Computationally identifying promoters has attracted scholars' attention for many years, and many encouraging results were obtained. However, it is still a challenging topic in bioinformatics.[17] In this work, we proposed a new feature extraction technique that combines PseKNC with PCSF for improving prediction *ACC*. A series of examinations demonstrated that our proposed method can distinguish promoter from non-promoter sequences with good performance. Thus, we established a predictor iProEP for providing convenience to scholars.

In the future work, many more promoters derived from other species will be collected for species-specific promoter prediction.[17,32] Moreover, although the combination of PseKNC and PCSF worked well in this study, new feature extraction techniques should be developed to further improve the performance of promoter prediction. Finally, with accumulation of more and more data and the development of a deep learning technique in many biological problems,[17,21,33–35] it is suitable to identify promoters by using a deep learning technique.

MATERIALS AND METHODS

Benchmark Dataset

A key step for constructing a powerful and robust prediction model is to construct an objective and strict benchmark dataset. In this work, we established five benchmark datasets including promoter and non-promoter sequences for five species (Table 5).

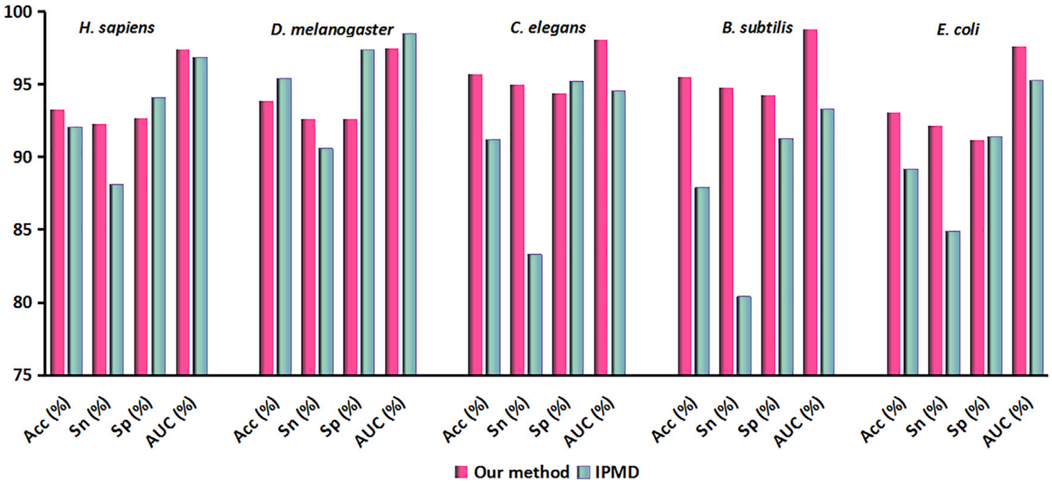


Figure 3. The Comparison between Our Proposed Method with IPMD Classifiers in 10-Fold Cross-Validation

Eukaryotic Promoter Database (EPD)[36] is a high-quality and non-redundant promoter resource and can be freely accessed at https://epd.epfl.ch//EPD_database.php. The 1,787 *H. sapiens* and 1,886 *D. melanogaster* Pol II promoter sequences were obtained from the EPD database. The 598 *C. elegans* promoter sequences were extracted from CEPDB (*C. elegans* promoter database; http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi). Each eukaryotic promoter is 300 bp long from 249 bp upstream to 50 bp downstream regions of TSS (TSS is regarded as 0-th site).

For prokaryote, 270 *B. subtilis* σ^{43} promoters were collected from DBTBS[37] (http://dbtbs.hgc.jp), and 741 *E. coli* K-12 σ^{70} promoter sequences were gained from RegulonDB[38] (http://regulondb.ccg.unam.mx/). All prokaryotic promoters have 81 nt with the region from −60 to +20 flanking TSS (TSS is regarded as the 0-th site).

The negative datasets were taken from the five species genome sequences. We randomly extracted 1,800 coding sequences and 1,800 introns from human DNA sequences from http://www.fruitfly.org/sequence/human-datasets.html [39] to generate the non-promoter dataset for *H. sapiens.* For *D. melanogaster*, a negative dataset including 2,859 coding sequences and 1,799 introns was downloaded from the website (http://www.fruitfly.org/sequence/drosophila-datasets.html).[40] The negative sample of *C. elegans* contains 600 coding sequences, and 600 introns were randomly extracted from Exon-Intron Database (EID).[41] For prokaryotes, all negative samples were randomly taken from the well-known database GenBank.[42] The number of non-promoter sequences for *B. subtilis* and *E. coli* are 600 (including 300 coding sequences and 300 convergent intergenic sequences) and 1,400 (including 700 coding sequences and 700 convergent intergenic sequences), respectively.

To get rid of the influence of noise data, we eliminated the sequences that contain other IUPAC code letters, such as "N," "S," and "W," from both positive and negative datasets. In order to ensure that the format of negative sequences can match the promoters, the lengths of eukaryotic and prokaryotic non-promoter sequences are also 300 and 81 bp, respectively. The details of the benchmark datasets were listed in Table 5.

It is well known that sequence similarity could influence the evaluation on the proposed mode.[43] We investigated the sequence similarity of the five species promoters by using CD-HIT. After setting the cutoff of sequence identity to 0.8 to exclude high similar promoters, we found that 98.0%, 99.3%, 95.0%, 98.5%, and 96.0% promoters for *H. sapiens*, *D. melanogaster*, *C. elegans*, *B. subtilis*, and *E. coli* remained, suggesting that the original datasets are objective enough to construct prediction models. Moreover, for the purpose of providing an objective comparison with the previous promoter prediction method IPMD, the same benchmarking datasets as used by IPMD are also provided. All data used in this study can be freely downloaded from http://lin-group.cn/server/iProEP/pages/download.php.

Pseudo *k*-Tuple Nucleotide Composition (PseKNC)

In general, the input of almost all the existing machine learning classification methods, such as SVM,[44–46] Random Forest,[47] and Artificial Neural Network,[48–50] must be a numeric value rather than a string sequence. Thus, each sample must be transferred into a fixed length of the feature vector.

A simple and common strategy to transform a DNA sample into a vector is to use its *k*-tuple nucleotide composition, which can be formulated by a vector $\boldsymbol{D}$ with 4^k elements according to the following formula:

$$\boldsymbol{D} = \left[f_1^{k-tuple} \; f_2^{k-tuple} \cdots f_i^{k-tuple} \cdots f_{4^k}^{k-tuple} \right]^{\boldsymbol{T}}, \qquad \text{(Equation 2)}$$

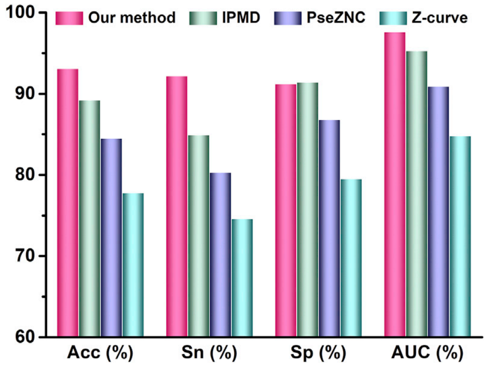


Figure 4. The Prediction Results of Four Methods on the Same *E. Coli* σ^{70} Promoter Data

where the symbol $\boldsymbol{T}$ means the transposition of a vector, and $f_i^{k-tuple}$ is the normalized frequency of the i-th k-tuple nucleotide component occurring in the DNA sequence.

In order to take both local and global sequence-order information of a DNA sequence into consideration, PseKNC[51,52] was proposed and has been widely utilized to represent DNA or RNA sequences.[53,54] Its basic principle is to combine the correlation of physiochemical properties of oligonucleotides and k-mer composition to formulate DNA sequences. There are two kinds of PseKNCs: type I and type II PseKNC. The former is also called the parallel correlation type, which mixes different physicochemical properties together to represent a nucleotide sequence with a vector containing $4^k + \Lambda$ components. The latter is named the series correlation type, which describes a nucleotide sequence by a vector containing $4^k + \lambda\Lambda$ factors. Comparing with the type I PseKNC, which has been widely and successfully applied in various bioinformatics fields,[8,55] few works focused on the application of type II PseKNC.[54,56] Considering the merit of type II PseKNC that different correlation information was separated independently, this work employed the type II PseKNC to transform sample sequences into vectors given as below:

$$\boldsymbol{D}_{pseKNC} = \left[d_1\ d_2\ \cdots\ d_{4^k}\ d_{4^k+1} \cdots d_{4^k+\lambda}\ d_{4^k+\lambda+1} \cdots d_{4^k\ +\lambda\Lambda}\right]^{\boldsymbol{T}},$$

(Equation 3)

where

$$d_u = \begin{cases} \dfrac{f_u^{k-tuple}}{\sum_{i=1}^{4^k} f_i^{k-tuple} + \omega \sum_{j=1}^{\lambda\Lambda} \tau_j}, & \left(1 \le u \le 4^k\right) \\ \dfrac{\omega \tau_{u-4^k}}{\sum_{i=1}^{4^k} f_i^{k-tuple} + \omega \sum_{j=1}^{\lambda\Lambda} \tau_j}, & \left(4^{k+1} \le u \le 4^{k+\lambda\Lambda}\right) \end{cases}.$$

(Equation 4)

$f_i^{k-tuple}$ has the same meaning as in Equation 2; λ is an integer number less than $L - k$, which reflects the correlation tiers or correlation rank along a DNA sequence; ω is a weight factor used to balance the effect of global correlation information and local property; and τ_j $(j = 1,\ 2,\ \cdots, \lambda\Lambda)$ represents the m-tier correlation factor, which describes the sequence-order correlation between all the m-tier contiguous k-tuple nucleotides along a DNA sequence. Here τ_j can be calculated by

Table 4. The Results for Cross-Species Examination

Kingdom	Model Training	Model Test	ACC (%)
Eukaryotes	*H. sapiens*	*D. melanogaster*	77.19
		C. elegans	66.63
	D. melanogaster	*H. sapiens*	68.41
		C. elegans	65.68
	C. elegans	*H. sapiens*	66.57
		D. melanogaster	69.58
Prokaryotes	*B. subtilis*	*E. coli*	75.95
	E. coli	*B. subtilis*	80.92

$$\begin{cases} \tau_1 = \dfrac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+1}^{1} \\ \tau_2 = \dfrac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+1}^{2} \\ \quad \cdots\cdots \\ \tau_\Lambda = \dfrac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+1}^{\Lambda} & \lambda < (L-\ k) \\ \quad \cdots\cdots \\ \tau_{\lambda\Lambda-1} = \dfrac{1}{L-k-\lambda+1} \sum_{i=1}^{L-\ k-\lambda+1} J_{i,i+1}^{\lambda\Lambda-1} \\ \tau_{\lambda\Lambda} = \dfrac{1}{L-k-\lambda+1} \sum_{i=1}^{L-k-\lambda+1} J_{i,i+1}^{\lambda\Lambda} \end{cases},$$

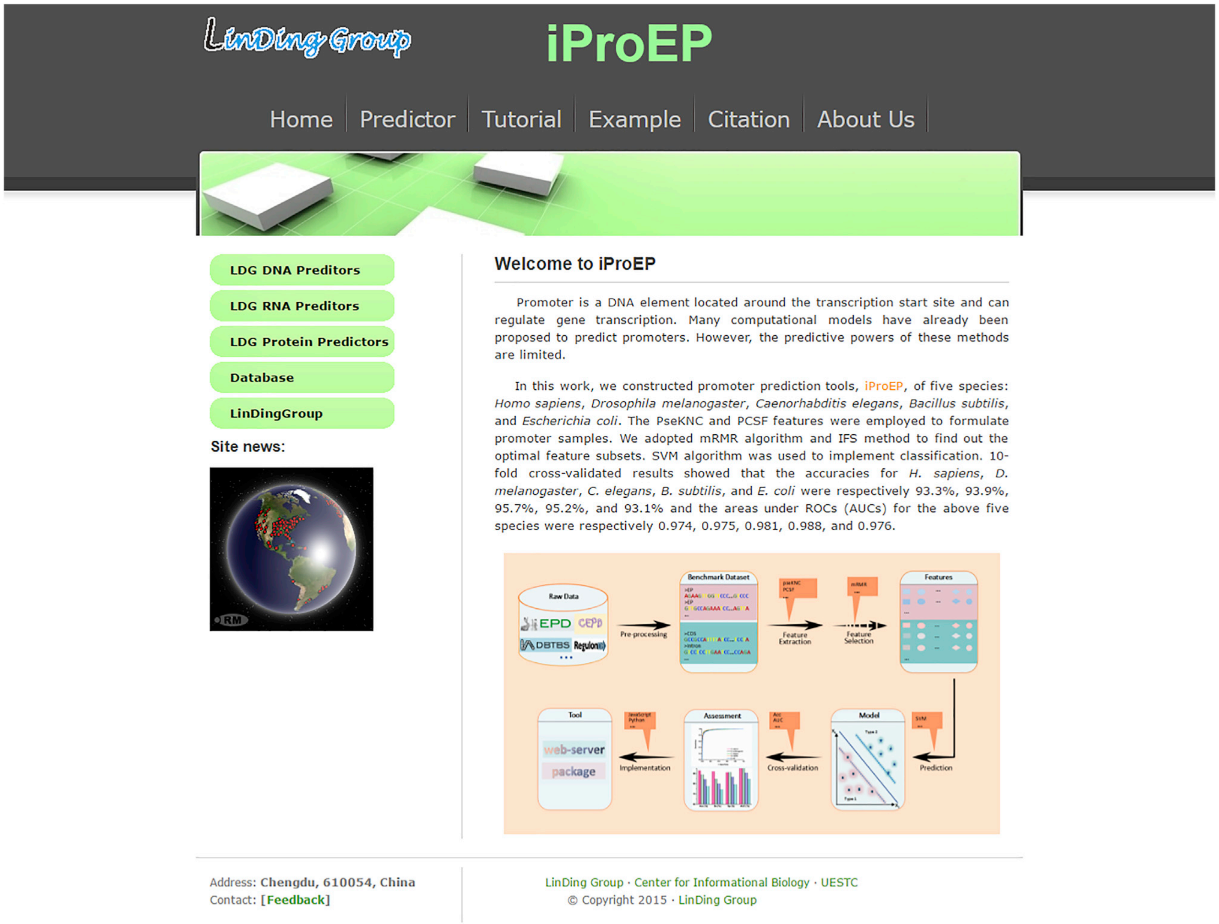
(Equation 5)

where

$$\begin{cases} J_{i,i+m}^{\xi} = H_\xi(R_iR_{i+1}) \bullet H_\xi(R_{i+m}R_{i+m+1}) \\ \xi = 1,\ 2,\ \bullet\bullet\bullet,\ \Lambda;\ m = 1,\ 2,\ \bullet\bullet\bullet,\ \lambda;\ i = 1,\ 2,\ \bullet\bullet\bullet,\ L-\lambda-1 \end{cases},$$

(Equation 6)

where $H_\xi(R_iR_{i+1})$ is a numerical value of the ξ-th physicochemical property for the dinucleotide R_iR_{i+1} at position i, $H_\xi(R_{i+m}R_{i+m+1})$ is the corresponding value for the dinucleotide $R_{i+m}R_{i+m+1}$ at position $i + m$, and Λ is the number of physicochemical properties. In this study, six DNA local structural properties of the 16 DNA dinucleotides were utilized in this work; the concrete values of three local translational properties (slide, shift, rise) and three local angular properties (roll, tilt, twist) were taken from Goñi et al.'s[57] work. It should be noted that the original values of six DNA local structural properties should be subjected to a standard



Welcome to iProEP

Promoter is a DNA element located around the transcription start site and can regulate gene transcription. Many computational models have already been proposed to predict promoters. However, the predictive powers of these methods are limited.

In this work, we constructed promoter prediction tools, iProEP, of five species: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Bacillus subtilis*, and *Escherichia coli*. The PseKNC and PCSF features were employed to formulate promoter samples. We adopted mRMR algorithm and IFS method to find out the optimal feature subsets. SVM algorithm was used to implement classification. 10-fold cross-validated results showed that the accuracies for *H. sapiens*, *D. melanogaster*, *C. elegans*, *B. subtilis*, and *E. coli* were respectively 93.3%, 93.9%, 95.7%, 95.2%, and 93.1% and the areas under ROCs (AUCs) for the above five species were respectively 0.974, 0.975, 0.981, 0.988, and 0.976.

Address: Chengdu, 610054, China
Contact: [Feedback]

LinDing Group · Center for Informational Biology · UESTC
© Copyright 2015 · LinDing Group

Figure 5. The Homepage of the iProEP Web Server
Available at http://lin-group.cn/server/iProEP/.

version by Equation 7 and then can be used in Equation 6 to calculate PseKNC:

$$H_{\xi}(R_iR_{i+1}) = \frac{H_{\xi}^{0}(R_iR_{i+1}) - \left\langle H_{\xi}^{0}(R_iR_{i+1})\right\rangle}{SD\left\langle H_{\xi}^{0}(R_iR_{i+1})\right\rangle}, \qquad \text{(Equation 7)}$$

where $H_{\xi}^{0}(R_iR_{i+1})$ is the original value of the ξ-th DNA local structural property for the dinucleotide R_iR_{i+1} at position i, the symbol<> means taking the average of the quantity therein for the 16 different combinations of A, G, C, T for R_iR_{i+1}, and *SD* represents the corresponding SD. The standard version of these physicochemical property values can be also found in many other DNA-related studies.[55] The superiority of the final standard 16 values converted by Equation 7 is that they will have a zero mean value over the 16 different dinucleotides and will not be changed if going through the same conversion procedure again.[58]

PCSF

By aligning promoter sequences for every species, we can construct a position-correlation scoring matrix (PCSM).[24,59] Each row in the PCSM consisted of factor p_{xi}, which is the probability of k-mer x at the i-th site of promoter samples. p_{xi} can be calculated by the following formula:

$$p_{xi} = \frac{n_{xi} + b_{xi}}{N_i + B_i}, \qquad \text{(Equation 8)}$$

where n_{xi} is the actual count of x appearing at the i-th site, and b_{xi} is the corresponding pseudocount. N_i indicates the sum of real counts of all k-mers at the i-th site (namely, positive sample number), and B_i is the corresponding sum of the pseudocount. If the sample size is not large enough, some k-mers will not be present when k increases. Hence, the pseudocount could improve

Table 5. The Detail Information of the Training Datasets for Five Species

			Non-promoter		
Kingdom	Species	Promoter	CDS	Non-CDS[a]	Location
Eukaryotes (300 bp)	*H. sapiens*	1,787	1,800	1,800	[−249, +50]
	D. melanogaster	1,886	1,799	2,859	[−249, +50]
Prokaryotes (81 bp)	*C. elegans*	598	600	600	[−249, +50]
	B. subtilis	270	300	300	[−60, +20]
	E. coli	741	700	700	[−60, +20]

CDS, coding sequences.
[a]Intron for eukaryotes and convergent intergenetic region for prokaryotes.

estimation of the probability p_{xi} for k-mer x at the i-th site. B_i and b_{xi} can be given by

$$\begin{cases} B_i = \sqrt{N_i} \\ b_{xi} = p_0\sqrt{N_i} \end{cases}, \qquad \text{(Equation 9)}$$

in which p_0 is the background frequency of k-mer, which is equal to $1/4^k$. With the increasing sample number N_i, the influence of pseudocounts will weaken, because of the slow increase of $\sqrt{N_i}$.

Some conservation sites of trimers for five species have been screened out by a great number of complex conservation analyses and *ACC* evaluations in Lin and Li.[24] Based on these sites and PCSM, the PCSF feature of positive and negative samples for five species can be expressed as

$$PCSF = \left[f_1\, f_2 \cdots f_i \cdots f_n\right], \qquad \text{(Equation 10)}$$

where n is the number of selected conservation sites, and each element is defined as

$$f_i = ln\left(p_{xi}/p_0\right). \qquad \text{(Equation 11)}$$

In this equation, p_0 is the background probability of each trimer ($p_0 = 1/4^3$), and p_{xi} can be obtained on the basis of PCSM.

mRMR

Commonly, picking out of the most useful features from the high-dimension data is a requisite step to exclude noise, improve prediction *ACC* and efficiency, avoid model overfitting, as well as build a robust model. In the present work, with the increase of two variables in Equation 4, k and λ, the dimension of PseKNC features will raise sharply, which may result in the curse of dimensionality. Therefore, it is absolutely necessary to find out the optimal features that could produce a robust model with highest *ACC*. mRMR is a popular feature selection technique that could calculate a score for each feature for measuring the importance of the feature.[60,61] It used a series of intuitive measures of relevance and redundancy to find a very compact subset from candidate features and has been widely used in data mining of biological processes.[62–65] For discrete features, two selection criteria, Mutual Information Difference criterion (MID) and Mutual Information Quotient criterion (MIQ), can be used to calculate the score of a feature. In the study, we chose the score from MIQ.

After scoring the PseKNC and PCSF features by mRMR, the IFS strategy with 5-fold cross-validation was applied to obtain the best feature subset that could produce the maximum prediction *ACC*. During the IFS procedure, the ranked features were added in the training set one by one according to mRMR rank; IFS strategy evaluates the performance of the top k-ranked features. The 5-fold cross-validation was to seek the best penalty coefficient c and width parameter γ for SVM models when obtaining the best feature subset.[54,56]

SVM

SVM is a widely employed machine learning algorithm based on statistical learning theory[66] and has been extended in bioinformatics fields.[67–73] The core idea of SVM is to seek out a classification hyperplane that can maximize the margin of the feature space. LibSVM is a popular softpackage for executing SVM[74] and can be freely downloaded from https://www.csie.ntu.edu.tw/∼cjlin/libsvm/. This study used LibSVM with radial basis function (RBF) to perform classification. We employed the grid search method with cross-validation to seek the best penalty coefficient c and width parameter γ. The searching space is as follows:

$$\begin{cases} c \in \left[2^{-5},\ 2^{15}\right], & step = 2 \\ \gamma \in \left[2^{-15},\ 2^{3}\right], & step = 2^{-1} \end{cases}. \qquad \text{(Equation 12)}$$

Performance Evaluation Metrics

In order to assess the quality of a predictor and compare different prediction tools, the following three indexes,[75] namely, the overall *ACC*, *Sn*, and *Sp*, were used and formulated as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(Equation 13)}$$

$$Sn = \frac{TP}{TP + FN} \qquad \text{(Equation 14)}$$

$$Sp = \frac{TN}{TN + FP}, \quad \text{(Equation 15)}$$

where *TP* (true positive) and *TN* (true negative) present the numbers of correctly identified promoters and non-promoters, respectively, and *FP* (false positive) and *FN* (false negative) denote the number of non-promoters incorrectly classified as promoters and the number of promoters incorrectly classified as non-promoters.

ROC analysis was used to measure the performance of the model with the varying of decision thresholds.[76]

AUTHOR CONTRIBUTIONS

H.D., W.C., and H.L. conceived and designed the study. H.-Y.L., Z.-Y.Z., Z.-D.S., and W.S. conducted the experiments. H.-Y.L. and Z.-Y.Z. implemented the algorithms. H.-Y.L., Z.-Y.Z., and Z.-D.S. established the web server. H.-Y.L., Z.-Y.Z., W.C., and H.L. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Scientific Foundation of China (grants 61772119 and 31771471), the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (no. C2017209244), and the Science Strength Promotion Programme of UESTC.

REFERENCES

1. Haberle, V., and Lenhard, B. (2016). Promoter architectures and developmental gene regulation. Semin. Cell Dev. Biol. *57*, 11–23.
2. Thomas, M.C., and Chiang, C.M. (2006). The general transcription machinery and general cofactors. Crit. Rev. Biochem. Mol. Biol. *41*, 105–178.
3. Slobodin, B., and Agami, R. (2015). Transcription initiation determines its end. Mol. Cell *57*, 205–206.
4. Pedersen, A.G., Baldi, P., Chauvin, Y., and Brunak, S. (1999). The biology of eukaryotic promoter prediction—a review. Comput. Chem. *23*, 191–207.
5. Hawley, D.K., and McClure, W.R. (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res. *11*, 2237–2255.
6. He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. BMC Syst. Biol. *12* (*Suppl 4*), 44.
7. Liang, Z.Y., Lai, H.Y., Yang, H., Zhang, C.J., Yang, H., Wei, H.H., Chen, X.X., Zhao, Y.W., Su, Z.D., Li, W.C., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics *33*, 467–469.
8. Lin, H., Deng, E.Z., Ding, H., Chen, W., and Chou, K.C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. *42*, 12961–12972.
9. Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., and Van de Peer, Y. (2008). Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res. *18*, 310–323.
10. Yang, J.Y., Zhou, Y., Yu, Z.G., Anh, V., and Zhou, L.Q. (2008). Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. BMC Bioinformatics *9*, 113.
11. Ohler, U. (2006). Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. Nucleic Acids Res. *34*, 5943–5950.
12. Davuluri, R.V., Grosse, I., and Zhang, M.Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet. *29*, 412–417.
13. Anwar, F., Baker, S.M., Jabid, T., Mehedi Hasan, M., Shoyaib, M., Khan, H., and Walshe, R. (2008). Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach. BMC Bioinformatics *9*, 414.
14. Burden, S., Lin, Y.X., and Zhang, R. (2005). Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences. Bioinformatics *21*, 601–607.
15. Gan, Y., Guan, J., and Zhou, S. (2009). A pattern-based nearest neighbor search approach for promoter prediction using DNA structural profiles. Bioinformatics *25*, 2006–2012.
16. Xu, W., Zhang, L., and Lu, Y. (2016). SD-MSAEs: Promoter recognition in human genome based on deep feature extraction. J. Biomed. Inform. *61*, 55–62.
17. Umarov, R.K., and Solovyev, V.V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS ONE *12*, e0171410.
18. Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian *N*^{6}-methyladenosine sites from mRNA. RNA *25*, 205–218.
19. Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of Deep Feature Representations and Handcrafted Features to Improve the Prediction of N^6-Methyladenosine Sites. Neurocomputing *324*, 3–9.
20. Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. Methods. Published online February 14, 2019. https://doi.org/10.1016/j.ymeth.2019.02.009.
21. Peng, L., Peng, M.M., Liao, B., Huang, G.H., Li, W.B., and Xie, D.F. (2018). The Advances and Challenges of Deep Learning Application in Biological Big Data Processing. Curr. Bioinform. *13*, 352–359.
22. Long, H.X., Wang, M., and Fu, H.Y. (2017). Deep Convolutional Neural Networks for Predicting Hydroxyproline in Proteins. Curr. Bioinform. *12*, 233–238.
23. Singh, S., Kaur, S., and Goel, N. (2015). A Review of Computational Intelligence Methods for Eukaryotic Promoter Prediction. Nucleosides Nucleotides Nucleic Acids *34*, 449–462.
24. Lin, H., and Li, Q.Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. Theory Biosci. *130*, 91–100.
25. Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. Nucleic Acids Res. *40*, 963–971.
26. Lin, H., Liang, Z.Y., Tang, H., and Chen, W. (2017). Identifying Sigma70 promoters with novel pseudo nucleotide composition (IEEE/ACM Trans. Comput. Biol. Bioinform). Published online February 8, 2017. https://doi.org/10.1109/TCBB.2017.2666141.
27. Liu, B., Yang, F., Huang, D.S., and Chou, K.C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.
28. Zhang, M., Li, F., Marquez-Lago, T.T., Leier, A., Fan, C., Kwoh, C.K., Chou, K.C., Song, J., and Jia, C. (2019). MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. Bioinformatics, 2019, btz016.
29. Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2018). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies (IEEE/ACM Trans. Comput. Biol. Bioinform). Published online March 15, 2018. https://doi.org/10.1109/TCBB.2018.2816032.
30. Yang, J., Chen, X., McDermaid, A., and Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. Bioinformatics *33*, 2586–2588.
31. Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., and Xu, Y. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. Nucleic Acids Res. *42*, W12–W19.

32. Shahmuradov, I.A., Umarov, R.K., and Solovyev, V.V. (2017). TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic Acids Res. *45*, e65.

33. Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., and Peng, S. (2018). Deep learning in omics: a survey and guideline. Brief. Funct. Genomics *18*, 41–57.

34. Yu, L., Sun, X., Tian, S.W., Shi, X.Y., and Yan, Y.L. (2018). Drug and Nondrug Classification Based on Deep Learning with Various Feature Selection Strategies. Curr. Bioinform. *13*, 253–259.

35. Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of Human Protein Subcellular Localization Using Deep Learning. J. Parallel Distrib. Comput. *117*, 212–217.

36. Dreos, R., Ambrosini, G., Cavin Périer, R., and Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. Nucleic Acids Res. *41*, D157–D164.

37. Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res. *36*, D93–D96.

38. Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J.S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J.A., et al. (2016). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. *44* (D1), D133–D143.

39. Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., Misra, S., and Rubin, G.M. (1999). The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics *153*, 135–177.

40. Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. (2002). Computational analysis of core promoters in the drosophila genome. Genome Biol *3*, RESEARCH0087.

41. Shepelev, V., and Fedorov, A. (2006). Advances in the Exon-Intron Database (EID). Brief. Bioinform. *7*, 178–185.

42. Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2015). GenBank. Nucleic Acids Res. *43*, D30–D35.

43. Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2019). Sequence Clustering in Bioinformatics: An Empirical Study. Brief. Bioinform., 2019, bby090.

44. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting Protein Structural Classes for Low-Similarity Sequences by Evaluating Different Features. Knowl. Base. Syst. *163*, 787–793.

45. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: A Sequence-Based Predictor for Identifying 2′-O-Methylation Sites in Homo sapiens. J. Comput. Biol. *25*, 1266–1277.

46. Li, D., Ju, Y., and Zou, Q. (2016). Protein Folds Prediction with Hierarchical Structured SVM. Curr. Proteomics *13*, 79–85.

47. Kandaswamy, K.K., Chou, K.C., Martinetz, T., Möller, S., Suganthan, P.N., Sridharan, S., and Pugalenthi, G. (2011). AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. *270*, 56–62.

48. Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules *22*, e1732.

49. Cao, R., Bhattacharya, D., Hou, J., and Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics *17*, 495.

50. Jiang, L., Zhang, J., Xuan, P., and Zou, Q. (2016). BP Neural Network Could Help Improve Pre-miRNA Identification in Various Species. BioMed Res. Int. *2016*, 9565689.

51. Chen, W., Lei, T.Y., Jin, D.C., Lin, H., and Chou, K.C. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. *456*, 53–60.

52. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.

53. Yu, C.Y., Li, X.X., Yang, H., Li, Y.H., Xue, W.W., Chen, Y.Z., Tao, L., and Zhu, F. (2018). Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate. Int. J. Mol. Sci. *19*, 183.

54. Dao, F.Y., Lv, H., Wang, F., Feng, C.Q., Ding, H., Chen, W., and Lin, H. (2019). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics *35*, 2075–2083.

55. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res. Int. *2014*, 623149.

56. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). Iterm-Pseknc: A Sequence-Based Tool for Predicting Bacterial Transcriptional Terminators. Bioinformatics *35*, 1469–1477.

57. Goñi, J.R., Pérez, A., Torrents, D., and Orozco, M. (2007). Determining promoter location based on DNA structure first-principles calculations. Genome Biol. *8*, R263.

58. Chou, K.C., and Shen, H.B. (2007). Recent progress in protein subcellular location prediction. Anal. Biochem. *370*, 1–16.

59. Li, Q.Z., and Lin, H. (2006). The recognition and prediction of sigma70 promoters in Escherichia coli K-12. J. Theor. Biol. *242*, 135–141.

60. Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. *27*, 1226–1238.

61. Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. Neurocomputing *173*, 346–354.

62. Kabir, M., Ahmad, S., Iqbal, M., and Hayat, M. (2019). iNR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. Genomics. Published online February 16, 2019. 10.1016/j.ygeno.2019.02.006.

63. Yuan, F., Lu, L., Zhang, Y., Wang, S., and Cai, Y.D. (2018). Data mining of the cancer-related lncRNAs GO terms and KEGG pathways by using mRMR method. Math. Biosci. *304*, 1–8.

64. Li, B.Q., Hu, L.L., Chen, L., Feng, K.Y., Cai, Y.D., and Chou, K.C. (2012). Prediction of protein domain with mRMR feature selection and analysis. PLoS ONE *7*, e39308.

65. Wang, S.P., Zhang, Q., Lu, J., and Cai, Y.D. (2018). Analysis and Prediction of Nitrated Tyrosine Sites with the Mrmr Method and Support Vector Machine Algorithm. Curr. Bioinform. *13*, 3–13.

66. Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. Mach. Learn. *20*, 273–297.

67. Manavalan, B., Shin, T.H., and Lee, G. (2018). PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front. Microbiol. *9*, 476.

68. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics, 2019, btz015.

69. Tang, H., Zhao, Y.W., Zou, P., Zhang, C.M., Chen, R., Huang, P., and Lin, H. (2018). HBPred: a tool to identify growth hormone-binding proteins. Int. J. Biol. Sci. *14*, 957–964.

70. Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., and Chou, K.C. (2019). Iprot-Sub: A Comprehensive Package for Accurately Mapping and Predicting Protease-Specific Substrates and Cleavage Sites. Brief. Bioinform. *20*, 638–658.

71. Manavalan, B., Shin, T.H., and Lee, G. (2017). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget *9*, 1944–1956.

72. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.

73. Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics *15*, 120.

74. Chang, C.C., and Lin, C.J. (2011). Libsvm: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. *2*, 27.

75. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief. Bioinform. bbz048.

76. Metz, C.E. (1978). Basic principles of ROC analysis. Semin. Nucl. Med. *8*, 283–298.