Computational and Structural Biotechnology Journal 18 (2020) 2445–2452

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

ncPro-ML: An integrated computational tool for identifying non-coding RNA promoters in multiple species

Qiang Tang [a], Fulei Nie [b,d], Juanjuan Kang [c], Wei Chen [a,b,d,*]

[a] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China
[b] Center for Genomics and Computational Biology, Scholl of Life Sciences, North China University of Science and Technology, Tangshan 063210, China
[c] Affiliated Foshan Maternity & Child Healthcare Hospital, Southern Medical University (Foshan Maternity & Child Healthcare Hospital), Foshan 528000, China
[d] School of Public Health, North China University of Science and Technology, Tangshan 063210, China

ARTICLE INFO

Article history:
Received 14 July 2020
Received in revised form 30 August 2020
Accepted 1 September 2020
Available online 10 September 2020

Keywords:
non-coding RNA
Promoter
Sequence length effect
Ensemble learning

ABSTRACT

The promoter is located near the transcription start sites and regulates transcription initiation of the gene. Accurate identification of promoters is essential for understanding the mechanism of gene regulation. Since experimental methods are costly and ineffective, developing efficient and accurate computational tools to identify promoters are necessary. Although a series of methods have been proposed for identifying promoters, none of them is able to identify the promoters of non-coding RNA (ncRNA). In the present work, a new method called ncPro-ML was proposed to identify the promoter of ncRNA in *Homo sapiens* and *Mus musculus*, in which different kinds of sequence encoding schemes were used to convert DNA sequences into feature vectors. To test the length effect, for each species, datasets including sequences with different lengths were built. The results demonstrated that ncPro-ML achieved the best performance based on the dataset with the sequence length of 221 nucleotides for human and mouse. The performances of ncPro-ML were also satisfying from both independent dataset test and cross-species test. The results indicate that the proposed predictor can server as a powerful tool for the discovery of ncRNA promoters. In addition, a web-server for ncPro-ML was developed, which can be freely accessed at http://www.bio-bigdata.cn/ncPro-ML/.

1. Introduction

Non-coding RNA (ncRNA) is a kind of transcripts that lack clear potential to encode proteins or peptides [1]. A large portion of the human genome is transcribed into ncRNA with many different forms, namely long-noncoding RNA (lncRNA), micro RNA (miRNA), circular RNA (circRNA), etc. [1–3]. Although ncRNAs lack potential to encode proteins, numerous investigations have shown that they play critical roles in many important biological processes including cell cycle, differentiation, development, metabolism, and so on [1,2,4–6]. Moreover, accumulated evidences have demonstrated that ncRNAs exhibit complex interactions with a broad spectrum of human diseases [1,2,7,8]. Deep sequencing of size-fractionated RNAs has become a primary technique for discovering ncRNAs, which generated a myriad of ncRNA candidates. However, the mechanisms of ncRNA are obscure or controversial in some biological process [9,10]. Therefore, in order to accurately understand their functions, the genomic annotations of the identified ncRNAs are necessary.

The first step of functional genomic annotation is promoter identification. The promoter is an important functional element in non-coding region, which immediately locates near and upstream of the transcription start site (TSS) and is mainly in charge of the gene transcription initiation. Due to their extensive roles in gene transcription, the accurate prediction of promoters becomes an essential step for understanding gene expression and the function of genetic regulatory networks. There were two main kinds of biological experiments for identification of promoters such as mutational analysis and immunoprecipitation assays [11–13]. Given that these methods were both expensive and time-consuming, computational methods have been proposed to identify promoters. In the past several years, several classifiers have been proposed to identify promoters in multiple species [14–16]. All these works concerned on the identification of promoters for coding genes. To the best of our knowledge, there exist

* Corresponding author: Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China.
E-mail address: greatchen@ncst.edu.cn (W. Chen).

https://doi.org/10.1016/j.csbj.2020.09.001

no computational methods able to identify promoters of non-coding RNA (ncRNA) genes.

Keeping this in mind, we proposed a support vector machine (SVM) based method, called ncPro-ML, to identify promoters of ncRNA. In order to comprehensively extract the sequence based information, eight kinds of feature representation schemes (binary and *k*-mer frequency [BKF], dinucleotide binary profile and frequency [DBPF], dinucleotide physical-chemical properties [DPCP], trinucleotide physical-chemical properties [TPCP], electron-ion interaction pseudopotentials of trinucleotide [triEIIP], ring-function-hydrogen-chemical properties [RFHCP], pseudo dinucleotide composition [PseDNC] and multivariate mutual information [MMI]) were used to convert DNA sequences into numerical vectors. To obtain a robust model, the feature selection process was utilized to select optimal feature subsets from the candidate feature list for each feature representation scheme. Based on multiple optimal subsets, we trained different models and integrated them by setting the weights according to the accuracy obtained from the five-fold cross-validation test. To demonstrate the effect of sequence length on predictive performance, distinct models based on different lengths ranging from 61 to 301 bp were tested as well. Finally, an easy-to-use webserver for ncPro-ML was developed, which is freely available at http://www.bio-bigdata.cn/ncPro-ML/. The flowchart on building ncPro-ML was shown in Fig. 1.

2. Materials and methods

2.1. Benchmark dataset

In this work, the promoter sequences of ncRNA from *Homo sapiens* and *Mus musculus* genome were obtained from the publicly available Eukaryotic Promoter Database (EPDnew) [17]. Compared with other TSS annotation databases, i.e. refTSS [18] and DBTSS/DBKERO [19], the EPD contains non-redundant collection promoters with stronger support from experimental data. To avoid the inclusion of noisy sequences, sequences which contains uncertain bases were removed. Considering that non-promoters do not have TSS, thus they were extracted from the downstream region of the promoter sequences. Thus, the dataset can be formulated as following,

$$\mathbb{S}_\xi = \mathbb{S}_\xi^+ \cup \mathbb{S}_\xi^- \tag{1}$$

where $\mathbb{S}_\xi^+$ is the positive dataset including promoter sequences. All these sequences are $\xi - bp$ long from (ξ-20) bp upstream to 20 bp downstream of the TSS (TSS is regarded at the 0th site). $\mathbb{S}_\xi^-$ is the negative dataset including non-promoter sequences. They are also $\xi - bp$ long, but start from 1000 bp downstream of the TSS. To demonstrate the effect of sequence length on predictive performance, a series of datasets based on different sequence lengths ranging from 61 to 221 bp with a step 20 bp, and 261 bp and 301 bp were built, which were formulated as following,

$$\mathbb{S}_\xi = \begin{cases} 61bp\ \xi = 61 \\ 81bp\ \xi = 81 \\ 101bp\ \xi = 101 \\ \cdots \\ 201bp\ \xi = 201 \\ 221bp\ \xi = 221 \\ 261bp\ \xi = 261 \\ 301bp\ \xi = 301 \end{cases} \tag{2}$$

The detail information about these datasets were given in Table 1. For the promoter and non-promoter sequences, 1170 and 1539 sequences of each length for human and mouse are used to train the model, and the rest are used as independent testing datasets to validate the performance of the model.

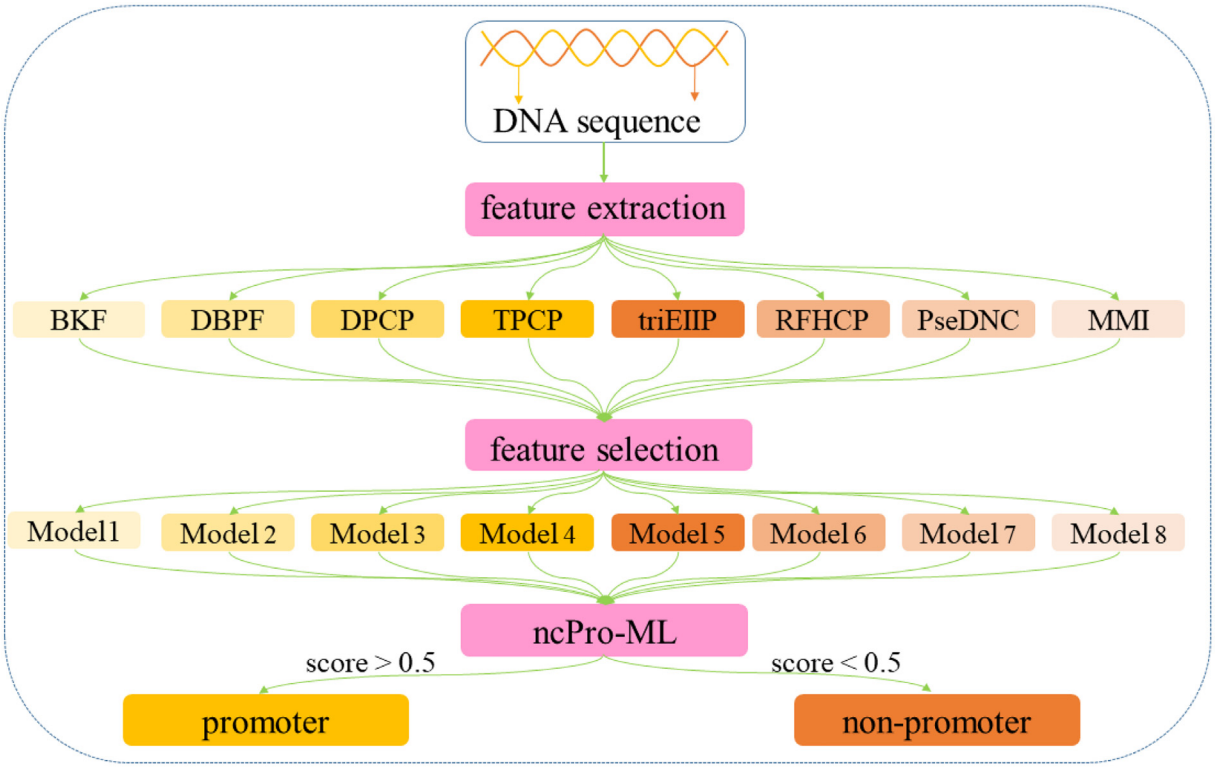

Fig. 1. The flowchart for building ncPro-ML.

Table 1
Detail information on the datasets used in this study.

Organism	Dataset Name	Promoter	number	Non-promoter	number
human	61 bp	$P_{40\text{-}20}$	2339	$N_{1000\text{-}1060}$	2339
	81 bp	$P_{60\text{-}20}$		$N_{1000\text{-}1080}$	
	101 bp	$P_{80\text{-}20}$		$N_{1000\text{-}1100}$	
	121 bp	$P_{100\text{-}20}$		$N_{1000\text{-}1120}$	
	141 bp	$P_{120\text{-}20}$		$N_{1000\text{-}1140}$	
	161 bp	$P_{140\text{-}20}$		$N_{1000\text{-}1160}$	
	181 bp	$P_{160\text{-}20}$		$N_{1000\text{-}1180}$	
	201 bp	$P_{180\text{-}20}$		$N_{1000\text{-}1200}$	
	221 bp	$P_{200\text{-}20}$		$N_{1000\text{-}1220}$	
	261 bp	$P_{240\text{-}20}$		$N_{1000\text{-}1260}$	
	301 bp	$P_{280\text{-}20}$		$N_{1000\text{-}1300}$	
mouse	61 bp	$P_{40\text{-}20}$	3077	$N_{1000\text{-}1060}$	3076
	81 bp	$P_{60\text{-}20}$		$N_{1000\text{-}1080}$	
	101 bp	$P_{80\text{-}20}$		$N_{1000\text{-}1100}$	
	121 bp	$P_{100\text{-}20}$		$N_{1000\text{-}1120}$	
	141 bp	$P_{120\text{-}20}$		$N_{1000\text{-}1140}$	
	161 bp	$P_{140\text{-}20}$		$N_{1000\text{-}1160}$	
	181 bp	$P_{160\text{-}20}$		$N_{1000\text{-}1180}$	
	201 bp	$P_{180\text{-}20}$		$N_{1000\text{-}1200}$	
	221 bp	$P_{200\text{-}20}$		$N_{1000\text{-}1220}$	
	261 bp	$P_{240\text{-}20}$		$N_{1000\text{-}1260}$	
	301 bp	$P_{280\text{-}20}$		$N_{1000\text{-}1300}$	

2.2. Feature representation algorithms

A given DNA sequence with L bp is defined as,

$$D = R_1R_2R_3R_4R_5R_6R_7 \cdots R_L \tag{3}$$

where $R_i \in \{A,C,G,T\}$ indicates the nucleotide at i-th position in the sequence. In this study, we utilized eight sequence-based feature representation algorithms to encode the sequences in the dataset.

2.2.1. Binary and k-mer frequency (BKF)

For the nucleotide binary profile, the nucleotides A, C, G and T are encoded by using the vectors (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) and (0, 0, 0, 1), respectively. Accordingly, a sequence can be represented by a 4L-dimensional feature vector. k-mer frequency is another way of representing DNA sequences, which refers to the frequency of all the possible k-tuple nucleotides in a given sequence. In this study, k was set to 2, 3 and 4. Thus, we could obtain three vectors with the dimension of 16, 64, 256, respectively. By combining the nucleotide binary profile and k-tuple nucleotide frequency, a sequence will be encoded by a (4L + 16 + 64 + 256) dimension vector.

2.2.2. Dinucleotide binary profile and frequency (DBPF)

Dinucleotide binary profile (DBP) and dinucleotide frequency were also widely used for sequence representation. For DBP, each dinucleotide type is encoded as a 4-dimensional vector containing 0 and 1. For instance, AA, AC, AG, were represented by (0, 0, 0, 0), (0, 0, 1, 0) and (0, 1, 0, 0), and so forth. The dinucleotide frequency is defined as following,

$$f_i = \frac{1}{|X_i|} C(R_{i-1}R_i),\ 2 \leqslant i \leqslant L \tag{4}$$

where $|X_i|$ is the length of sub-sequence $(R_1R_2 \ldots R_i)$ in the sequence D, and $C(R_{i-1}R_i)$ is the occurrence frequency of the dinucleotide $R_{i-1}R_i$ in the X_i-length sub-sequence. Therefore, for a given sequence, the dimension of the vector based on DBPF is 4×(L-1) + L-1).

2.2.3. Dinucleotide physical-chemical properties (DPCP)

Physicochemical properties are also important information for genomic functional elements identifications and were incorporated into promoter prediction[20,21]. Inspired by those works, 15 different physicochemical properties, namely PC1, F-roll; PC2, F-tilt; PC3, F-twist; PC4, F-slide; PC5, F-shift; PC6, F-rise; PC7, roll; PC8, tilt; PC9, twist; PC10, slide; PC11, shift; PC12, rise; PC13, energy; PC14, enthalpy; and PC15, entropy, were employed to encode sequences in the dataset. The values of the 15 properties for each dinucleotide were provided in Supplementary Table S1. Since the values of different properties vary greatly, their original values were normalized to the range of [0, 1] by using the max–min normalization method. The DPCP is formulated as following,

$$DPCP(i) = f(i) \times PC(X_i) \tag{5}$$

where i is one of the 16 dinucleotides, $f(i)$ is the frequency of the i-th dinucleotide in a sequence and the X represents one of the 15 physicochemical properties. Based on DPCP, a given sequence can be encoded as a 240 (16 × 15)-dimensional vector.

2.2.4. Trinucleotide physical-chemical properties (TPCP)

Similar to DPCP, the following 11 physical-chemical properties: PC1, bendability (DNase); PC2, bendability (consensus); PC3, trinucleotide GC content; PC4, nucleosome positioning; PC5, consensus (roll); PC6, consensus (rigid); PC7, DNase I (rigid); PC8, molecular weight (daltons); PC9, nucleosome (rigid); PC10, nucleosome; and PC11, DNase I were used to define TPCP. The values of these 11 physicochemical properties for each trinucleotide are listed in Supplementary Table S2. These values were normalized as described above before performing the following calculation. The TPCP is formulated as

$$TPCP(i) = f(i) \times PC(X_i) \tag{6}$$

where X is one of the 11 physicochemical properties, i is one of the trinucleotides and $f(i)$ is the frequency of the i-th trinucleotide in a sequence. Then, a sequence can be encoded as a 704 (64 × 11) -dimensional vector.

2.2.5. Electron-ion interaction pseudopotentials of trinucleotide (triEIIP)

The EIIP was an effective feature encoding method which has been widely used bioinformatics [22–24]. The EIIP values of the four nucleotides are given in Supplementary Table S3. The composition of each sequence can be represented as a 64-dimensional feature vector E as follows,

$$E = [EIIP_{AAA} \cdot f_{AAA}, EIIP_{AAC} \cdot f_{AAC}, \cdots, EIIP_{TTT} \cdot f_{TTT}] \quad (7)$$

where $EIIP_{xyz} = EIIP_x + EIIP_y + EIIP_y + EIIP_z$, is the EIIP value of the nucleotide *xyz*, and $x, y, z \in \{A, C, G, T\}$, f_{xyz} is the frequency of trinucleotide *xyz* in the sequence.

2.2.6. Ring-function-hydrogen-chemical properties (RFHCP)

The deoxyribonucleic acid is composed of four nucleic acids that have different chemical properties in terms of ring structures, strength of hydrogen bonds and chemical functionality [25]. Considering the number of rings, A and G are grouped together because they both contains two rings and the others are in one group which only have one ring. In terms of hydrogen bond, C and G can be distributed in the same group since they form strong hydrogen bonds, whereas A and T form weak hydrogen bonds and thus belong to the other group. In the aspect of the chemical functionality, A and C can be classified into the amino group, while G and T can be classified into the keto group. Accordingly, three coordinates (*x*, *y*, *z*) were used to represent the chemical properties of the four nucleotides. The *x* , *y* and *z* stand for the ring structure, the hydrogen bond and the chemical functionality, respectively. Each nucleotide *i* in the sequence can be encoded by (x_i, y_i, z_i), where

$$x_i = \begin{cases} 1 \ if R_i \in \{A, G\} \\ 0 \ if R_i \in \{C, T\} \end{cases}, \ y_i = \begin{cases} 1 \ if R_i \in \{A, T\} \\ 0 \ if R_i \in \{C, G\} \end{cases}, \ z_i = \begin{cases} 1 \ if R_i \in \{A, C\} \\ 0 \ if R_i \in \{G, T\} \end{cases} \quad (8)$$

Moreover, the density d_i of a nucleotide at position *i* was defined as following [26],

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{L} f(R_j), \ f(R_j) = \begin{cases} 1 & if R_j = q \\ 0 & other cases \end{cases} \quad (9)$$

where $|N_i|$ is the length of subsequence $(R_1 R_2 \ldots R_i)$ in the sequence D. By integrating the two schemes, a sequence can be encoded by a 4 × L- dimensional vector.

2.2.7. Pseudo dinucleotide composition (PseDNC)

PseDNC can reflect both short-range and long-range sequence-order information by calculating the dinucleotide nucleotide composition and the correlation of physics-chemical properties from a consider sequence [27]. In this study, we used six types of local structural parameters (Slide, Shift, Rise, Twist, Tilt and Roll) to characterize the spatial arrangements of any two successive base pairs. For a given sequence, it can be denoted as a $16 + \lambda$ dimension vector formulated as following,

$$D = [d_1 d_2 \ldots d_{16} d_{16+1} \ldots d_{16+\lambda-1} d_{16+\lambda}] \quad (10)$$

where

$$\begin{cases} d_u = \frac{f_u}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} \ (1 \leqslant u \leqslant 16) \\ d_u = \frac{\omega \theta_{u-16}}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} \ (16 + 1 \leqslant u \leqslant 16 + \lambda) \end{cases} \quad (11)$$

where $f_u (u = 1, 2, \ldots, 16)$ is the normalized frequency of the *u*-th k-tuple nucleotide composition, ω is the weight factor range from 0.1 to 1 with a step 0.1 and λ is the number of the total counted ranks or tiers of the correlations along a DNA sequence. In this study, we set a search strategy for λ ranges from 1 to 10. The *j*-th tire structural correlation factor θ_j that reflects the local structure correlation between all the *j*-th most contiguous dinucleotide along a DNA sequence and can be given by

$$\theta_j = \frac{1}{L - j - 1} \sum_{i=1}^{L-j-1} \Theta(R_i R_{i+1}; R_{i+j} R_{i+j+1}) \ (j = 1, 2, \ldots, \lambda; \lambda < L) \quad (12)$$

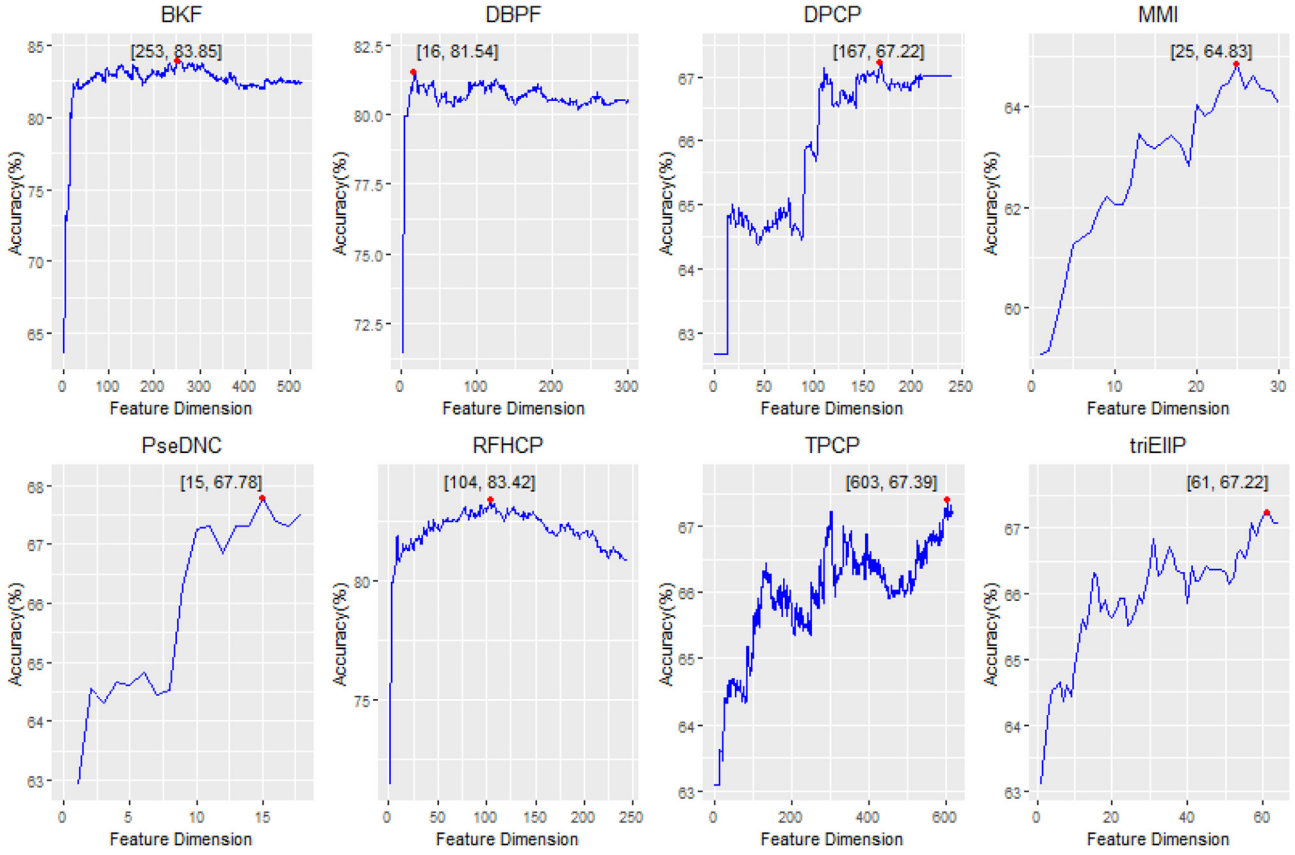

Fig. 2. The variation of Acc versus the increment of feature dimension for identifying human ncRNA promoters based on the 61 bp dataset.

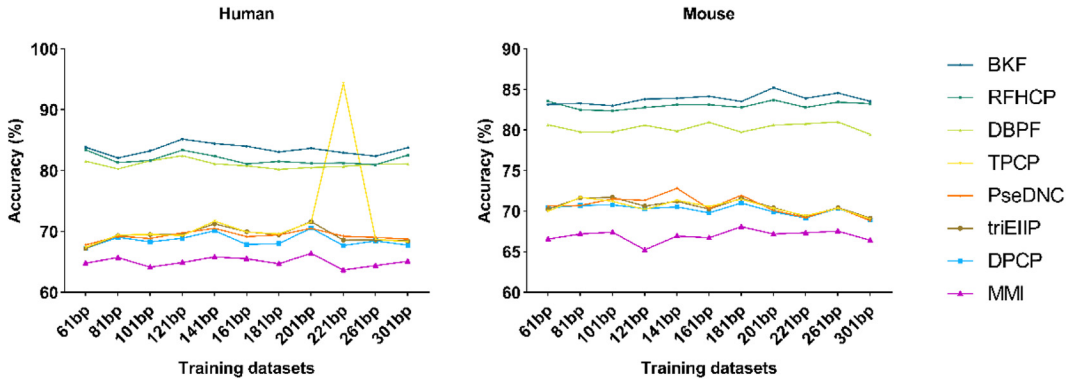

Fig. 3. The accuracy of models based on different features and datasets in human and mouse.

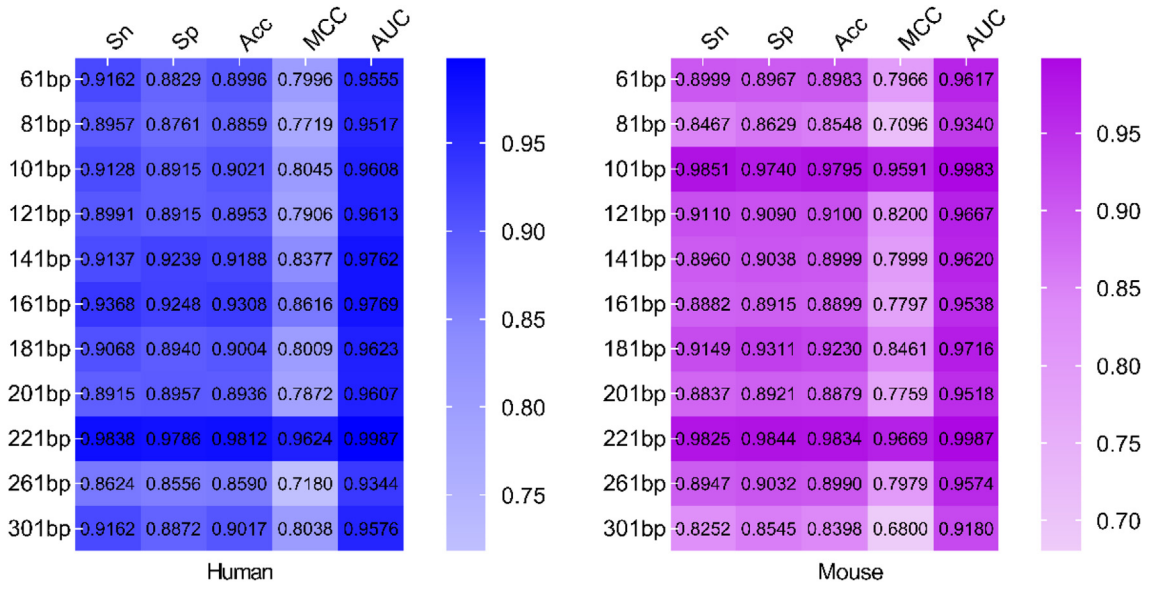

Fig. 4. The performance of predictors based on different datasets. The performance was measured in term of Sn, Sp, Acc, MCC and AUC.

$\Theta(R_iR_{i+1};R_{i+j}R_{i+j+1})$ is the correlation function and can be defined by

$$\Theta(R_iR_{i+1};R_{i+j}R_{i+j+1}) = \frac{1}{6}\sum_{v=1}^{6}\left[P_v(R_iR_{i+1}) - P_v(R_{i+1}R_{i+j+1})\right]^2 \quad (13)$$

where the $P_v(r_iR_{i+1})$ is the value of the v-th DNA local structural property for the dinucleotide R_iR_{i+1} at position i in the sequence.

2.2.8. Multivariate mutual information (MMI)

The feature encoding method of multivariate mutual information (MMI) was proposed by Pan et al. and has been widely used in the field of bioinformatics [23,28]. In order to use MMI on a DNA sequence, we first define 2-tuple nucleotide composition set T_2 and 3-tuple nucleotide composition set T_3 as follows.

$$\begin{cases} T_2 = \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\} \\ T_3 = \{\ AAA,\ AAC,\ AAG,\ AAT,\ ACC,\ ACG,\ ACT,\ AGG,\ AGT,\ ATT, \\ CCC, CCG,\ CCT,\ CGG,\ CGT,\ CTT,\ GGG,\ GGT,\ GTT,\ TTT\} \end{cases} \quad (14)$$

According to the formula described by Pan et al.[28], the MMI can be defined as follows:

$$\begin{cases} I(R_iR_j) = f(R_iR_j)ln\frac{f(R_iR_j)}{f(R_i)f(R_j)} \\ I(R_iR_jR_k) = I(R_iR_j) + \frac{f(R_iR_k)}{f(R_k)}ln\frac{f(R_iR_k)}{f(R_k)} - \frac{f(R_iR_jR_k)}{f(R_jR_k)}ln\frac{f(R_iR_jR_k)}{f(R_jR_k)} \end{cases} \quad (15)$$

where $f(R_i)$ is the frequency of R_i in the sequence, the $f(R_iR_j)$ is the frequency of categories R_iR_j appearing in the T_2 feature on a sequence and $f(R_iR_jR_k)$ frequency of categories $R_iR_jR_k$ appearing in the T_3 feature on a sequence. Accordingly, a sequence is represented by 10 + 20 = 30 features generated according to Eq. (14).

2.3. Feature selection

Feature selection is a key step to find the most useful features to improve the classification accuracy and reduce the number of features. For eliminating redundant and irrelevant features, we first applied the F-score method to calculate the importance of features and yielded a feature ranking list regarding their classification importance. And then, we used the sequential forward search (SFS) strategy to find the optimal feature representations [29,30]. For the strategy of SFS, features from the ranked feature list was added one by one from higher to lower rank to select the sub-features. Then, the SVM based models were trained and tested

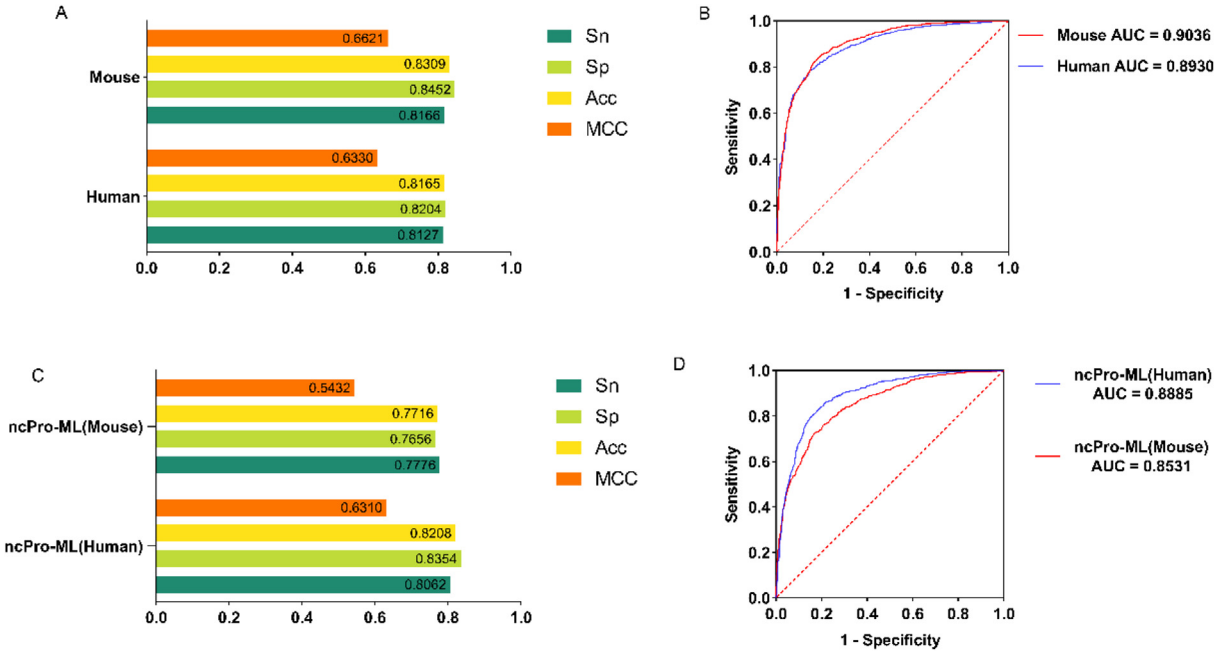

Fig. 5. Performance of ncPro-ML based on independent datasets (A and B) and cross-species datasets (C and D). In A and B, the Mouse and Human represent the model human and mosue in ncPro-ML, respectively. In C and D, the ncPro-ML(Human) denote using human model in ncPro-ML to perform the mouse independent testing datasets, and vice versa.

based on the sub-features by using a 5-fold cross-validation. Finally, the sub-features with the best performance was recognized as the optimal feature set.

2.4. Building promoter recognition models based on SVM

SVM is a powerful supervised-learning algorithm based on the statistical learning theory and has widely applied to handle many biological problems, such as recognizing special peptides [31–33] and protein [34], disease diagnosis [35]. In this study, the LIBSVM package [36] was used to train the SVM and built a model that could discriminate between ncRNA promoter and non-promoter sequence, and the most commonly used Radial Basis Function (RBF) was selected as its kernel function. To achieve the optimal performance, we optimized the SVM using a grid search approach to filter the regularization parameter C and kernel parameter γ. The search ranges for both of the parameters are given as following,

$$\begin{cases} 2^{-5} \leqslant C \leqslant 2^{15}, \ with step size of 2 \\ 2^{-15} \leqslant \gamma \leqslant 2^{-5}, \ with step size of -2 \end{cases} \tag{16}$$

2.5. Performance measures

The performance of the proposed method was evaluated by using four commonly used metrics, namely sensitivity (Sn), specificity (Sp), accuracy (Acc) and the Mathew's correlation coefficient (MCC). They are calculated as follows:

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+TN+FN+FP} \\ MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \end{cases} \tag{17}$$

In equation (17), TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively.

Besides, the receiver operating characteristic curve (ROC) was also employed to evaluate the overall performance of the proposed model. ROC is an objective metric that designed to simultaneously display the true positive rate against the false positive rate at every possible classification threshold and has been widely used in diverse fields. The value of the area under ROC curve (AUC), which ranges from 0.5 to 1, can reduce the ROC performance to a single scalar value representing expected performance. The higher the value of the AUC the better performance is implied.

3. Results and discussion

3.1. Feature optimization

In this study, we generated eleven feature representations by using eight kinds of feature encoding schemes that represents sequence information in different sides. For some feature encoding schemes, the longer the sequences, the greater the dimension of the feature vector. Take BKF as an example, the feature dimension was $4 \times L + 16 + 64 + 256$, a total 1540-D features will be generated when the sequence with 301 nucleotides. Such a problem may lead to an increase in classifiers training time and a reduction of their predictive performance. To address these issues, we conducted a 5-fold cross-validation test for each feature representation scheme based on optimal features obtained by using the feature selection strategy. To intuitively analyze the results, in Fig. 2, we plotted the variation of Acc versus the increment of feature dimension for identifying human ncRNA promoters based on dataset $\boldsymbol{S}_{61}$. The red point in the figure is the highest Acc for each feature representations. It was found that the maximum Acc of 83.85% was achieved when 253 BKF derived optimal sub-features were used. This result demonstrates that the feature dimension is greatly reduced and the accuracy of the model is significantly improved by using the feature optimization strategy. The results of feature selection process for human and mouse based on different datasets were shown in the Supplementary Figure S1 to Figure S21.

3.2. Construct ncPro-ML by integrating multiple models

Multiple model integration method is an important pattern classification technique to obtain better performance and can avoid the potential deviation generated by a single classifier [37]. Therefore, we combined these eight models according to the weighted sum of their prediction scores, where the weights were normalized by the Acc of a single model divided by the sum of the Accs of the eight models. For example, based on the human dataset $\mathbb{S}_{81}$, the weight of 0.14 $(82.09/(82.09+80.30+69.06+65.73+69.23+81.32+69.40+69.32))$ was obtained for the model based on feature BKF. Similarly, the weight of 0.1369, 0.1178, 0.1121, 0.1181, 0.1387, 0.1183 and 0.1182 were obtained for the models based on feature DBPF, DPCP, MMI, PseDNC, RFHCP, TPCP and triEIIP, respectively. For the final model, the prediction score was the sum of the prediction score of the eight models based on their weights. Finally, eight integrated models were constructed based on different lengths datasets for human and mouse. The weights of the eight models in different lengths datasets for human and mouse were listed in Supplementary Table S4 and Table S5.

3.3. Effect of sequences length on model performance

We have built eleven datasets including different sequence length ranging from 61 to 301 nucleotides for human and mouse. The best accuracy produced by the feature selection process for each feature representation method of different datasets were shown in Fig. 3. The number of features for BKF, RFHCP and DBPF were larger than others, and increased with the length of the training sequences. As it can be seen in the Fig. 3, those models built based on BKF, RFHCP and DBPF obtained better predictive performance than the models based on other kinds of features.

Although the variation of the performance based on different datasets are not significant for the eight kinds of features in both human and mouse, the best predictive accuracy was obtained by using BKF based on dataset $\mathbb{S}_{121}$ for human and based on dataset $\mathbb{S}_{201}$ for mouse. Especially, the TPCP has a very high predictive accuracy for human on dataset $\mathbb{S}_{221}$. Taken together, the performance of eight models remained relatively stable for all datasets with different lengths.

According to the weights of each model based on different datasets, we constructed eight integrated predictors by adding weights to a model for human and mouse, respectively. Due to the consistency among the accuracy of different datasets for each model, we evaluated the performance of the eight predictors by using self-test to determine the best sequence length for human and mouse. Where the self-test refers to using the training datasets to validate the constructed model. The results obtained from the experiments to verify the impact of the sequence length variation on the predictors performance are shown in Fig. 4. For human and mouse, the eight predictors trained based on different datasets all yielded a better predictive performance.

From these results, we chose the predictor training based on the $\mathbb{S}_{221}$ as the final predictor for human and mouse. The two predictors obtained the highest Acc of 98.12% and 98.34%, and were used to build ncPro-ML for identifying ncRNA promoters in human and mouse, respectively. Moreover, we compared the performance of SVM with that of different machine learning based methods, namely Naive Bayes, Random Forest, Random Tree, Logistic, k-nearest neighbor (KNN) and SVM based on the datasets $\boldsymbol{S}_{221}$. The results from five-fold cross validation test demonstrated that the SVM based method yielded the best performance in term of Acc (Figure S22).

3.4. Performance assessment of ncPro-ML based on the independent datasets

To assess the generalization ability and robustness of the predictor, ncPro-ML was validated on the independent datasets. The performances of the predictor based on the independent datasets $\mathbb{S}_{221}$ for identifying promoters in human and mouse were shown in Fig. 5A and B. The predictor achieved the accuracy of 81.65% with the sensitivity of 81.27%, specificity of 82.04% and MCC of 0.633 for human, and accuracy 83.09% with the sensitivity of 81.66%, specificity of 84.52% and MCC of 0.6621 for mouse. The corresponding AUC is 0.8930 and 0.9036 for human and mouse, respectively. These results indicate that the proposed method is reliable for identifying ncRNA promoters in human and mouse.

To demonstrate generalization ability of ncPro-ML, the cross-species validation was also performed. Accordingly, the model trained in one species (human or mouse) was tested on the independent datasets of the other species. The predictive results were shown in Fig. 5 C and D.

The human and mouse specific model achieved the Acc of 82.08% and 77.16% to identify promoters in mouse and human independent datasets, respectively. The corresponding AUCs were 0.8885 and 0.8531. The excellent performance of ncPro-ML indicates that the proposed predictor can server as a powerful tool for the discovery of new ncRNA promoters.

4. Conclusion

Accurate identification of promoters is essential for understanding the mechanism of the gene regulation process and is also a fundamental step for functional annotation of a new genome. Therefore, numerous computational approaches have been proposed by using different machine learning methods. However, to the best of our knowledge, there is no predictor specifically for identifying the ncRNA promoters. To address this challenge, we proposed the first machine-learning based method ncPro-ML to identify ncRNA promoters in human and mouse. In order to make ncPro-ML yield excellent performance, for both human and mouse, eleven datasets composed of sequences with different lengths were constructed to evaluate the sequence length required for training a predictor with the best performance. The performance of ncPro-ML on independent datasets indicate that ncPro-ML is good enough for identify ncRNA promoters in human and mouse. In addition, results from cross-species evaluation demonstrate that ncPro-ML have the ability to identify ncRNA promoters in other species as well. For the convenience of scientific community, a user-friendly web server for ncPro-ML was provided at http://www.bio-bigdata.cn/ncPro-ML/. We hope it could become a useful tool for identifying ncRNA promoters.

Funding

This work was supported by the National Nature Scientific Foundation of China (No. 31771471), the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), and the Xinglin Scholar Research Premotion Project of Chengdu University of TCM (NO. ZRQN2019015).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.09.001.

References

[1] Matsui M, Corey DR. Non-coding RNAs as drug targets. Nat Rev Drug Discov 2017;16:167–79.
[2] Zhang W, Zhang H, Yang H, et al. Computational resources associating diseases with genotypes, phenotypes and exposures. Brief Bioinform 2019;20:2098–115.
[3] Kimura T. Metal-mediated epigenetic regulation of gene expression. Yakugaku Zasshi 2017;137:273–9.
[4] Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature 2016;539:452–5.
[5] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–97.
[6] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009;136:215–33.
[7] Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell 2009;136:629–41.
[8] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet 2009;10:155–9.
[9] Wang K, Chang H. Molecular Mechanisms of Long Noncoding RNAs. Mol Cell 2011;43:904–14.
[10] Wong C-M, Tsang F-C, Ng I-L. Non-coding RNAs in hepatocellular carcinoma: molecular functions and pathological implications. Nat Rev Gastroenterol Hepatol 2018;15:137–51.
[11] Matsumine H, Yamamura Y, Hattori N, Kobayashi T, Kitada T, Yoritaka A, Mizuno Y. A Microdeletion of D6S305 in a Family of Autosomal Recessive Juvenile Parkinsonism (PARK2). Genomics 1998;49:143–6.
[12] Kim J-W, Zeller KI, Wang Y, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV. Evaluation of Myc E-Box Phylogenetic Footprints in Glycolytic Genes by Chromatin Immunoprecipitation Assays. MCB 2004;24:5923–36.
[13] Dahl JA, Collas P. A rapid micro chromatin immunoprecipitation assay (microChIP). Nat Protoc 2008;3:1032–45.
[14] Oubounyt M, Louadi Z, Tayara H, et al. DeePromoter: Robust Promoter Predictor Using Deep Learning. Front Genet 2019;10:286.
[15] Wang S, Cheng X, Li Y, Wu M, Zhao Y. Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns. Sci Rep 2018;8:17695.
[16] Lin H, Deng EZ, Ding H, et al. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 2014;42:12961–72.
[17] Meylan P, Dreos R, Ambrosini G, et al. EPD in 2020: enhanced data visualization and extension to ncRNA promoters. Nucleic Acids Res 2020;48:D65–9.
[18] Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, Kasukawa T. refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. J Mol Biol 2019;431:2407–22.
[19] Suzuki A, Kawano S, Mitsuyama T, et al. DBTSS/DBKERO for integrated analysis of transcriptional regulation. Nucleic Acids Res 2018;46:D229–38.
[20] Brick K, Watanabe J, Pizzi E. Core promoters are predicted by their distinct physicochemical properties in the genome of Plasmodium falciparum. Genome Biol 2008;9:R178.
[21] 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.
[22] Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation 2006;1:197–202.
[23] Wei L, Su R, Luan S, et al. Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics 2019;35:4930–7.
[24] He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics 2019;35:593–601.
[25] Chen W, Yang H, Feng P, et al. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017;33:3518–23.
[26] Chen W, Feng P, Song X, Lv H, Lin H. iRNA-m7G: Identifying N7-methylguanosine Sites by Fusing Multiple Features. Mol Ther Nucleic Acids 2019;18:269–74.
[27] Chen W, Lin H, Chou K-C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 2015;11:2620–34.
[28] Pan G, Jiang L, Tang J, et al. A Novel Computational Method for Detecting DNA Methylation Sites with DNA Sequence Information and Physicochemical Properties. Int J Mol Sci 2018;19.
[29] Ru B, 'T Hoen PAC, Nie F, et al. PhD7FASTER: predicting clones propagating faster from the Ph.D.-7 phage display peptide library. J Bioinform Comput Biol 2014;12:1450005.
[30] Liu K, Chen W. iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics 2020;36:3336–42.
[31] Tang Q, Nie F, Kang J, Ding H, Zhou P, Huang J. NIEluter: Predicting peptides eluted from HLA class I molecules. J Immunol Methods 2015;422:22–7.
[32] He B, Kang J, Ru B, Ding H, Zhou P, Huang J. SABinder: A Web Service for Predicting Streptavidin-Binding Peptides. Biomed Res Int 2016;2016:1–8.
[33] Li N, Kang J, Jiang L, He B, Lin H, Huang J. PSBinder: A Web Service for Predicting Polystyrene Surface-Binding Peptides. Biomed Res Int 2017;2017:1–5.
[34] Kang J, Fang Y, Yao P, Li N, Tang Q, Huang J. NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition. Interdiscip Sci Comput Life Sci 2019;11:108–14.
[35] Kang J, Yu S, Lu S, Xu G, Zhu J, Yan Na, Luo D, Xu K, Zhang Z, Huang J. Use of a 6-miRNA panel to distinguish lymphoma from reactive lymphoid hyperplasia. Sig Transduct Target Ther 2020;5.
[36] Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2(3):1–27.
[37] Tang Q, Kang J, Yuan J, et al. DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species. Bioinformatics 2020;36:3327–35.