Bioinformatics, 34(1), 2018, 33–40
doi: 10.1093/bioinformatics/btx579
Advance Access Publication Date: 14 September 2017
Original Paper

OXFORD

Sequence analysis

iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC

Bin Liu[1,2,*], Fan Yang[1], De-Shuang Huang[3,*] and Kuo-Chen Chou[2,4,5,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China, [2]The Gordon Life Science Institute, Boston, MA 02478, USA, [3]Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China and [5]Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 13, 2017; revised on August 27, 2017; editorial decision on September 11, 2017; accepted on September 13, 2017

Abstract

Motivation: Being responsible for initiating transaction of a particular gene in genome, promoter is a short region of DNA. Promoters have various types with different functions. Owing to their importance in biological process, it is highly desired to develop computational tools for timely identifying promoters and their types. Such a challenge has become particularly critical and urgent in facing the avalanche of DNA sequences discovered in the postgenomic age. Although some prediction methods were developed, they can only be used to discriminate a specific type of promoters from non-promoters. None of them has the ability to identify the types of promoters. This is due to the facts that different types of promoters may share quite similar consensus sequence pattern, and that the promoters of same type may have considerably different consensus sequences.

Results: To overcome such difficulty, using the multi-window-based PseKNC (pseudo K-tuple nucleotide composition) approach to incorporate the short-, middle-, and long-range sequence information, we have developed a two-layer seamless predictor named as 'iPromoter-2 L'. The first layer serves to identify a query DNA sequence as a promoter or non-promoter, and the second layer to predict which of the following six types the identified promoter belongs to: σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54} and σ^{70}.

Availability and implementation: For the convenience of most experimental scientists, a user-friendly and publicly accessible web-server for the powerful new predictor has been established at http://bioinformatics.hitsz.edu.cn/iPromoter-2L/. It is anticipated that iPromoter-2 L will become a very useful high throughput tool for genome analysis.

Contact: bliu@hit.edu.cn or dshuang@tongji.edu.cn or kcchou@gordonlifescience.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Promoters are short consensus sequences of DNA fragments at specific location, whose role is to initiate transcription of a particular gene. Bacterial promoters consist of a purine at the transcription start site (TSS), the hexamer TATAAT centered at −10, and another hexamer centered at −35 (Feklistov *et al.*, 2014). In *Escherichia coli*, there are several σ factors in the RNA polymerase, each of which causes the RNA polymerase to initiate at a set of promoters

defined by the specific sequence positions of -35 and -10, and has the function of identifying the promoter and facilitating the binding of the RNA polymerase to the promoter. Each σ factor has a certain function. For instance, σ^{24} and σ^{32} are associated with the heat-shock response, σ^{28} is involved in the expression of flagellar genes during normal growth, σ^{54} is used for nitrogen metabolism, and σ^{70} is responsible for transcription of most genes under normal conditions (Janga and Collado-Vides, 2007; Potvin *et al.*, 2008).

The types of promoters are defined by how the σ factors identify the promoters. However, individual promoter usually differs from the consensus at one or more positions, and hence making it difficult to identify promoter through traditional experimental methods. Nevertheless, the evolutionary conservation of the consensus still plays a key role in promoter identification. Several computational methods have been proposed to overcome these difficulties. For example, based on the sequence-level conservation of hexamer segments, Li and Lin (2006) proposed a position-correlation scoring function (PCSF) to identify σ^{70}-promoter sequences. Introducing the variable window technique via the regular Z-curve method (Zhang, 1997; Zhang *et al.*, 2003), Song (2012) proposed an approach called 'variable-window Z-curve' to identify promoters. Two years later, by using DNA duplex stability as discriminative characteristic, an interesting method (de Avila e Silva *et al.*, 2014) was developed to identify σ^{54}- and σ^{28}-promoters. Meanwhile, based on the PseKNC (Chen *et al.*, 2014c), an extended version of PseAAC (Chou, 2001) for dealing with DNA/RNA sequences, Lin *et al.* (2014) proposed a predictor called 'iPro54-PseKNC' to identify σ^{54}-promoters.

All the aforementioned methods did play admirable roles in stimulating the development of this important field, but they were designed for identifying a specific type of promoter. In other words, each of them is actually a binary classifier. In this study, we are to develop a two-layer predictor: its first layer can be used to identify a DNA sequence as a promotor or non-promoter; its second layer used to identify which type of the identified promoter is. In the last decade or so, various two-layer predictors such as MemType-2 L (Chou and Shen, 2007a), NR-2 L (Wang *et al.*, 2011), GPCR-2 L (Xiao *et al.*, 2011a), Quat-2 L (Xiao *et al.*, 2011b), iAMP-2 L (Xiao *et al.*, 2013), iEnhancers-2 L (Liu *et al.*, 2016a) and 2 L-piRNA (Liu *et al.*, 2017d) were developed for identifying different types or functions of membrane proteins, nuclear receptors, G-protein-coupled receptors, quaternary proteins, antimicrobial peptides, enhancers and piwi-interacting RNAs among many others (see, e.g. Chen *et al.*, 2012; Han *et al.*, 2014). To our best knowledge, however, so far there is no two-layer promoter predictor whatsoever that can be used to identify promoters and their types as well. This study was initiated in an attempt to fill such an empty area.

2 Materials and methods

2.1 Benchmark dataset

According to Chou's five-step rules (Chou, 2011) and complied by a series of recent investigators (see, e.g., (Chen *et al.*, 2016b; Feng *et al.*, 2017; Jia *et al.*, 2016d; Liu *et al.*, 2015a, 2017a,d,e; Xu *et al.*, 2013b, 2014a), to develop a useful statistical predictor for a biological system, the first important thing is to construct or select a good benchmark dataset to train the model and an independent benchmark dataset to test it. But as elucidated in (Chou and Shen, 2007b), if the model is to be tested by the jackknife or K-fold cross-validation (where K represents any integer) tests (Chou and Zhang, 1995), there is no need at all to artificially separate a benchmark dataset into two: one for training and one for testing. In other words, one good benchmark dataset would suffice. Why? Because the result obtained by the jackknife or K-fold cross-validation test is actually a combination from an array of independent dataset tests (Chou and Shen, 2007b).

To construct a high-quality benchmark dataset, all the promoter samples in this study were collected from the experiment-confirmed ones (each with 81 bp) as collected in RegulonDB (version 9.3) (Gama-Castro *et al.*, 2016). Furthermore, to remove redundancy and reduce homologous bias, the CD-HIT software (Li and Godzik, 2006) was used to ensure that none of the samples included has $\geq$ 0.8 pairwise sequence identity to any other in a same type of promoters.

The non-promoter sequences were randomly extracted from the middle regions of long coding sequences and convergent intergenic regions in *E.coli* K-12 genome (Lin *et al.*, 2014, 2017). The non-promoter samples are also of 81 bp, with the same length as of the promoter samples. Likewise, the same cutoff procedure was used to ensure that none of the non-promoters included has $\geq$0.8 pairwise sequence identity to each other.

For clarity, the benchmark dataset $\mathbb{S}$ thus obtained is formulated by

$$\begin{cases} S = \mathbb{S}^+ \cup \mathbb{S}^- \\ \mathbb{S}^+ = \mathbb{S}^+(\sigma^{24}) \cup \mathbb{S}^+(\sigma^{28}) \cup \mathbb{S}^+(\sigma^{32}) \cup \mathbb{S}^+(\sigma^{38}) \cup \mathbb{S}^+(\sigma^{54}) \cup \mathbb{S}^+(\sigma^{70}) \end{cases} \quad (1)$$

where $\mathbb{S}^+$ is the positive set containing 2, 860 promoter samples, while $\mathbb{S}^-$ is the negative set also containing 2, 860 non-promoter samples. The symbol $\cup$ denotes the 'union' in the set theory. As we can see from Equation (1), $\mathbb{S}^+$ is a union of (or formed by) six positive subsets. They are $\mathbb{S}^+(\sigma^{24})$, $\mathbb{S}^+(\sigma^{28})$, $\mathbb{S}^+(\sigma^{32})$, $\mathbb{S}^+(\sigma^{38})$, $\mathbb{S}^+(\sigma^{54})$, and $\mathbb{S}^+(\sigma^{70})$, containing 484 promoter samples of σ^{24}, 134 of σ^{28}, 291 of σ^{32}, 163 of σ^{38}, 94 of σ^{54} and 1694 of σ^{70}, respectively.

The codes of all these samples as well as their detailed sequences are given in Supplementary Material S1.

2.2 Sample formulation with multi-window-based PseKNC

With the avalanche of biological sequences occurring in the post-genomic age, one of the most challenging problems in computational biology is how to formulate them with a discrete model but meanwhile being able to keep some useful sequence order or pattern information. To address this problem, the concept of PseAAC (Chou, 2001) was proposed. Since then, various different modes PseAAC were developed (see, e.g. Ahmad *et al.*, 2016; Esmaeili *et al.*, 2010; Hajisharifi *et al.*, 2014; Kumar *et al.*, 2015; Liu *et al.*, 2013; Meher *et al.*, 2017; Mohabatkar *et al.*, 2011; Nanni and Lumini, 2008; Rahimi *et al.*, 2017) and a long list of references cited in Chou, 2009). Encouraged by the successes of PseAAC for protein/peptide sequences, its extended version called PseKNC (pseudo K-tuple nucleotide composition) was also proposed (Chen *et al.*, 2014c, 2015c; Liu *et al.*, 2015c, 2016b) to deal with DNA/RNA sequences (Chen *et al.*, 2015b) for various problems in genome analyses (see, e.g. Chen *et al.*, 2013; Kabir and Hayat, 2016; Liu *et al.*, 2017a; Qiu *et al.*, 2014) as well as a long list of references cited in two recent papers (Chou, 2017; Liu *et al.*, 2017b).

In this study, we are to introduce a different mode of PseKNC, the so-called 'multi-window-based pseudo K-tuple nucleotide composition' or 'multi-window-based PseKNC', which is actually an extension of the flexible-sliding-window approach used to predict

signal peptides and their cleavage sites (Chou and Shen, 2007c; Shen and Chou, 2007). The concrete procedures are as follows.

Each of the DNA samples in the benchmark dataset [cf. Equation (1) and Supplementary Material S1] can be expressed as

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_{81} \ (i = 1,\ 2,\ 3,\ \cdots,\ 81) \tag{2}$$

where

$$\mathrm{N}_i \in \{A\ (\textbf{adenine}),\ C\ (\textbf{cytosine}),\ G\ (\textbf{guanine}),\ T\ (\textbf{thymine})\} \tag{3}$$

Suppose a sliding-window (Chou and Shen, 2007c; Shen and Chou, 2007) is expressed by $[\xi,\ \delta]$, where ξ denotes the width of the window and δ its step. For the DNA sample as given in Equation (2), the total number of the fragments generated by sliding $[\xi,\ \delta]$ along its sequence is given by

$$\eta = \ \mathrm{INT}\left[\frac{L-\xi+\delta}{\delta}\right] \tag{4}$$

where the symbol INT denotes an "integer-cutting operator" meaning to take the integer part of the quantity within the brackets right after it, $L = 81$ is the length of the original sample. For example, substituting $L = 81$, $\xi\ = 8$ and $\delta = \ 3$ into Equation (4), we obtain $\eta = \ 25$. In other words, using the sliding window of [8, 3] on the DNA sample of Equation (2), we can obtain an array of 54 DNA fragments.

For each of the 25 sub-sequences, we can define a PseKNC vector given by (Chen *et al.*, 2014c)

$$D(\mathrm{sub}) = [\phi_1\ \ \phi_2\ \ \cdots\ \ \phi_{\mathrm{u}}\ \ \cdots\ \ \phi_{\mathrm{Z}}]^{\mathrm{T}} \tag{5}$$

where (Chen *et al.*, 2015b)

$$\phi_{\mathrm{u}} = \begin{cases} \dfrac{f_u^{K-\mathrm{tuple}}}{\sum_{i=1}^{4^{\mathrm{K}}} f_{\mathrm{u}}^{\mathrm{K-tuple}} + \mathrm{w}\sum_{j=1}^{\lambda}\theta_j} & \left(1 \le \mathrm{u} \le 4^{\mathrm{K}}\right) \\ \dfrac{\mathrm{w}\theta_{\mathrm{u}-4^{\mathrm{K}}}}{\sum_{i=1}^{4^{\mathrm{K}}} f_{\mathrm{u}}^{\mathrm{K-tuple}} + \mathrm{w}\sum_{j=1}^{\lambda}\theta_j} & \left(4^{\mathrm{K}}+1 \le \mathrm{u} \le 4^{\mathrm{K}}+\lambda = \mathrm{Z}\right) \end{cases} \tag{6}$$

In the above equation, $f_u^{\mathrm{K-tuple}}$ is the uth component of the K-tuple nucleotide composition for the sub-DNA sample; the parameter w denotes the weight factor; λ represents the highest counted rank of the correlation along the DNA sub-sequence (Liu *et al.*, 2015c); and

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta(\mathrm{N}_i\mathrm{N}_{i+1}, \mathrm{N}_{i+\mathrm{j}}\mathrm{N}_{i+j+1})\ \ (j = 1,\ 2, \ldots, \lambda) \tag{7}$$

where the correlation factor $\Theta(N_iN_{i+1}, N_{i+j}N_{i+j+1})$ is given by

$$\Theta\left(\mathrm{N}_i\mathrm{N}_{i+1}, \mathrm{N}_{i+j}\mathrm{N}_{i+j+1}\right) = \frac{1}{\Phi}\sum_{v=1}^{\Phi}\left[P_v(\mathrm{N}_i\mathrm{N}_{i+1}) - P_v\left(\mathrm{N}_{i+j}\mathrm{N}_{i+j+1}\right)\right]^2 \tag{8}$$

where Φ is the number of physicochemical properties considered; $P_v(\mathrm{R}_i\mathrm{R}_{i+1})$ is the numerical value of the v-th physicochemical property for the dinucleotide at position i. In this study, six physicochemical properties (Rise, Roll, Shift, Slide, Tilt and Twist) for the DNA dinucleotides are used. There are a total of $4\times\ \ 4\times 6 = \ \ 96$ physicochemical property values, which can be obtained from (Chen *et al.*, 2013). But note that when substituting the original physicochemical property values into Eq.8, they were subjected to a standard conversion according to the following equation (Chou and Shen, 2006)

$$P_v(\mathrm{N_iN}_{i+1}) \Leftarrow \frac{P_v(\mathrm{N_iN}_{i+1}) - \ < P_v >}{\mathrm{SD}(P_v)} \tag{9}$$

where the $P_v(N_iN_{i+1})$ is the original value of the v-th DNA physicochemical property of the dinucleotide N_iN_{i+1}at position i; the symbol $\langle\ \rangle$ means the average value of the quantity therein for $4\times 4 = \ \ 16$ different dinucleotides, and SD denotes the corresponding standard deviation. The advantage to carry out the standard conversion is that the converted values obtained by Equation (9) will have a zero mean value over the 16 different dinucleotides, and will remain unchanged if they go through the same conversion procedure again (Chou and Shen, 2007b). The 96 physicochemical indexes obtained via Equation (9) can be found in the paper (Chen *et al.*, 2013) as well.

As we can see from above, the sub-sequence **D**(sub) of Equation (5) is well defined by a vector with $4^K + \lambda$ pseudo components. Since there are η sub-sequences [cf. Equation (4)] for a sample of Equation (2), by concatenating the corresponding η vectors, the sample is finally defined by a vector of $\eta\times(4^K+\lambda)$ pseudo components.

2.3 Optimizing imbalanced training datasets

As we can see from the text under Equation (1) as well as Supporting Information S1, the six subsets in $\mathbb{S}^+$ are very imbalanced to one another. For instance, the largest subset $\mathbb{S}^+(\sigma^{70})$ contains 1694 samples but the smallest subset $\mathbb{S}^+(\sigma^{54})$ contains only 94 samples. The former is over 18 times of the latter. Although this might reflect the real world where the number of σ^{54} promoters is much less than that of σ^{70}promoters, a predictor trained by such a highly skewed dataset would inevitably lead to a bias consequence: many σ^{54} promoters might be incorrectly predicted as σ^{70} (Jia *et al.*, 2016c; Liu *et al.*, 2015d; Xiao *et al.*, 2015). Therefore, it is important to find an effective approach to optimize the imbalanced training dataset and minimize this kind of bias consequence. To realize this, we took the following procedures in this study.

We used the IHTS (inserting hypothetical training samples) treatment to add some corresponding hypothetical or theoretical promoter samples into the subsets that are smaller than the largest one, making each of them contain exactly the same number of samples as the largest one. The details of how to generate the theoretical promoters for each promoter type have been elaborated in (Jia *et al.*, 2016d) and hence there is no need to repeat here. By doing so, we can generate five new subsets that have the same size as $\mathbb{S}^+(\sigma^{70})$; i.e.

$$\begin{aligned}\|\mathbb{S}^+(\sigma^{24}|bal)\| = \|\mathbb{S}^+(\sigma^{28}|bal)\| = \|\mathbb{S}^+(\sigma^{32}|bal)\| = \|\mathbb{S}^+(\sigma^{38}|bal)\| \\ = \|\mathbb{S}^+(\sigma^{54}|bal)\| = \|\mathbb{S}^+(\sigma^{70})\| = 1\ \ 694\end{aligned} \tag{10}$$

where the symbol $\|\ \|$ means taking the total number of the samples for the set therein.

Note that the theoretical samples generated via the IHTS treatment can only be expressed by their feature vectors as defined in Equation (5), but not the real DNA segment samples as given in Supplementary Material S1. To do so, however, is perfectly fine since the data directly used to train a predictor are actually the samples' feature vectors but not their sequence codes. This is the key to optimize the training dataset by the size-balancing approach, and its rationale of such an approach will be further elucidated later.

2.4 Random forest

Random forest (RF) is one of the most popular and powerful machine learning classifiers. It has been widely used in many fields of computational biology (see, e.g. Jia *et al.*, 2015, 2016a,b,e; Kandaswamy *et al.*, 2011; Lin *et al.*, 2011; Liu *et al.*, 2016c; Pugalenthi *et al.*, 2012; Qiu *et al.*, 2016a, 2017b; Xu *et al.*, 2013a).

The formulation of RF and the detailed procedures in using it have been clearly described in (Breiman, 2001), and hence there is no need to repeat here. In this study, the Scikit-learn (Pedregosa *et al.*, 2011) was used as the implementation of the RF algorithm with command line 'RandomForestClassifier(criterion='gini', max_features= 'sqrt', min_samples_split = 2, min_samples_leaf =1, $\mathscr{F}$ = optimized value)', where the optimized values of the parameter $\mathscr{F}$ (the number of trees in the forest) were 1900, and 800 for the first layer and second layer, respectively.

2.5 The predictor's workflow

The predictor established via the aforementioned procedures is named as 'iPromoter-2 L', where 'iPromoter' stands for 'identification of promoters', and '2 L' for 'two layers'. The first layer serves to predict whether a query DNA sequence sample is of promoter or not, while the second layer is used to further identify which type the identified promoter belongs to. The model in the first layer was trained by the dataset $\mathbb{S}$ of Equation (1), while the model in the second layer trained by the balanced datasets in Equation (10). Illustrated in Figure 1 is a flowchart to show how the two seamless models are working.

2.6 Target cross-validation

In this study, the 5-fold cross-validation (Chou and Zhang, 1995) was adopted to evaluate prediction quality. When conducting the cross-validation for the second-layer model, however, some special consideration is needed. This is because after the ITHTS treatments (Jia *et al.*, 2016d), except for the dataset $\mathbb{S}^+(\sigma^{70})$, all the other five ones contain many theoretical samples. For example, the subset $\mathbb{S}^+(\sigma^{24}|\text{bal})$ contains 1694 − 484 = 1210 σ^{24} theoretical promoters, the subset $\mathbb{S}^+(\sigma^{28}|\text{bal})$ contains 1694 − 134 = 1560 σ^{28} theoretical promoters, and so forth. It would be perfectly rational to use such a datasets to train a predictor, but not for validation. This is because the validation should be carried out against all the experiment-confirmed samples in the benchmark dataset but not the theoretical samples added later. Such validation is called 'target cross-validation', meaning the validation is targeting on experiment-confirmed data only rather than theoretical ones. As for how to carry out the target cross validation, refer to the papers (Jia *et al.*, 2016c,d), where a detailed step-by-step guide was provided, and hence there is no need to repeat here.

2.7 Performance evaluation metrics

To provide a set of more intuitive and easier-to-understand scales for measuring the prediction quality, the Chou's criterion used in predicting signal peptides was adopted in this study. According to that criterion, the prediction quality for the promoters and their types can be defined by the following four metrics (Feng *et al.*, 2013)

$$\begin{cases} Sn(i) = 1 - \dfrac{N_-^+(i)}{N^+(i)} & 0 \le Sn \le 1 \\ Sp(i) = 1 - \dfrac{N_+^-(i)}{N^-(i)} & 0 \le Sp \le 1 \\ Acc(i) = 1 - \dfrac{N_-^+(i) + N_+^-(i)}{N^+(i) + N^-(i)} & 0 \le Acc \le 1 \\ MCC(i) = \dfrac{1 - \left(\dfrac{N_-^+(i)}{N^+(i)} + \dfrac{N_+^-(i)}{N^-(i)}\right)}{\sqrt{\left(1 + \dfrac{N_+^-(i) - N_-^+(i)}{N^+(i)}\right)\left(1 + \dfrac{N_-^+(i) - N_+^-(i)}{N^-(i)}\right)}} & -1 \le MCC \le 1 \end{cases} \tag{11}$$

where $i = 1, 2, \cdots, M$, and Sn, Sp, Acc, and MCC represent the sensitivity, specificity, accuracy, and Mathew's correlation coefficient, respectively (Chen *et al.*, 2007), M is the total number of classes considered for the system, and i denotes the i-th class or type therein.

Thus, for the first-layer prediction, we have $M = 2$ and $i = 1$ or 2 for promoter or non-promoter, respectively.

For the second-layer prediction, we have $M = 6$ and $i = 1, 2, 3, 4, 5,$ or 6 for σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54}, or σ^{70} promoters, respectively.

In the metrics of Equation (11), $N^+(i)$ is the total number of the samples investigated in the i-th class, whereas $N_-^+(i)$ is the number of the samples in $N^+(i)$ that are incorrectly predicted to be of other class; $N^-(i)$ is the total number of samples in any class but not the ith class, whereas $N_+^-(i)$ is the number of the samples in $N^-(i)$ that are incorrectly predicted to be of the ith class.

Owing to their intuitive merit, the metrics of Equation (11) have been widely used to study many different problems in genome and proteome analyses (see, e.g. (Chen *et al.*, 2014a, 2015a; Ding *et al.*, 2014; Jia *et al.*, 2016e,f; Liu *et al.*, 2015b, 2016a,b,c,d; Qiu *et al.*, 2016b).

It is instructive to point out, however, the set of equations defined in Equation (11) is valid only for the single-label systems in which each sample belongs to one and only one class. For the multi-label systems where some samples may simultaneously belong to two or more classes, a completely different set of metrics as defined in (Chou, 2013) is needed. The multi-label systems have been increasingly found in system biology (Chou *et al.*, 2011, 2012; Cheng *et al.*, 2017a; Wu *et al.*, 2011) and system medicine (Cheng *et al.*, 2017b,c,d,e,f; Qiu *et al.*, 2016b; Xiao *et al.*, 2013).

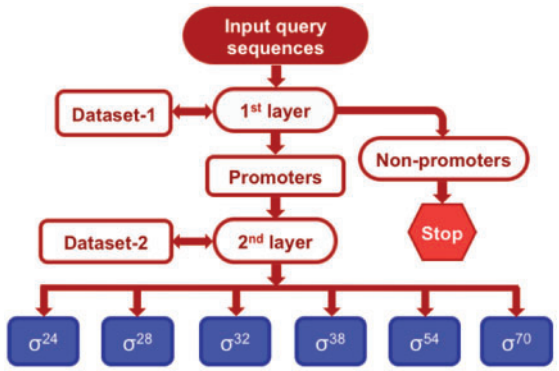

Fig. 1. A flow chart to show how the iPromoter-2 L is working, where Dataset-1 means the first sub-equation of Equation (1), and Dataset-2 means the dataset formed by the subsets in Equation (10)

3 Results and discussion

3.1 Determination of parameter

It can be seen from Sections 2.2 and 2.4 that the feature vector used to define the DNA sample contains five parameters ξ, δ, K, λ, and w, and that the RF algorithm used to run the prediction contains a parameter $\mathscr{F}$. In this study, their ranges can be formulated as follows

$$\begin{cases} 5 \le \xi \le 20 \\ 1 \le \delta \le \xi \\ 3 \le K \le 5 \\ 2 \le \lambda \le \xi - K + 1 \\ 0 \le w \le 1 \\ 600 \le F \le 2000 \end{cases} \tag{12}$$

Table 1. The prediction quality of iPromoter-2 L as measured by the four metrics of Equation (11) and examined by the target cross-validation on the experiment-confirmed data of Supplementary Material S1

		Metrics				
		Sn[c]	Sp[c]	Acc[c]	MCC[c]	Test method
First layer[a]		79.20%	84.16%	81.68%	0.6343	5-fold cross
		78.92%	84.37%	81.64%	0.6338	Jackknife
Second layer[b]	σ^{24}-promoter	72.52%	96.93%	93.50%	0.7338	5-fold cross
	σ^{28}-promoter	42.54%	99.49%	96.82%	0.5708	5-fold cross
	σ^{32}-promoter	52.58%	99.14%	94.41%	0.6524	5-fold cross
	σ^{38}-promoter	15.34%	99.48%	94.69%	0.2962	5-fold cross
	σ^{54}-promoter	53.19%	99.57%	94.04%	0.6459	5-fold cross
	σ^{70}-promoter	95.34%	59.35%	80.66%	0.6056	5-fold cross

[a]The parameters used for the first-layer sub-predictor are $\xi = 8$, $\delta = 3$, K = 3, $\lambda = 2$, $w = 0.5$ and $\mathscr{F} = 1900$.
[b]The parameters used in the second-layer sub-predictor are $\xi = 8$, $\delta = 3$, K = 4, $\lambda = 2$, $w = 0.5$ and $\mathscr{F} = 800$.
[c]See Equation (11) for the definition of the metrics.

And their final values are given in Table 1 that were determined via the procedures of optimizing the prediction quality as done in (Jia *et al.*, 2015, 2016b,e; Qiu *et al.*, 2016b, 2017b; Xiao *et al.*, 2016). As we can see from Table 1, the metrics scores for the 1st-layer predictor obtained by the 5-fold cross-validation test are nearly the same as those by the jackknife test. Accordingly, to reduce the computational time in this study, only the 5-fold cross-validations were adopted to estimate the metrics scores for the secondnd-layer predictor.

3.2 Comparison with the existing methods

Listed in Table 2 are the success rates reported by the four state-of-the-art methods (de Avila e Silva *et al.*, 2014; Li and Lin, 2006; Lin *et al.*, 2014; Song, 2012) in identifying the promoters. As we can see from the table, the scores achieved by the iPrometer-2 L predictor proposed in this paper are higher than its cohorts in all the four metrics. Particularly, the current predictor can go one step deeper to discriminate the identified promoter among the six different types indicated in Equatio (1). This is beyond the reach of the other four predictors.

Table 2. Comparison of the proposed predictor with the four state-of-the-art methods for identifying promoters with 5-fold cross validation

Method	Sn (%)[a]	Sp(%)[a]	Acc (%)[a]	MCC[a]	Capacity[g]
PCSF[b]	78.92	70.70	74.81	0.4980	No
vw Z-curve[c]	77.76	82.80	80.28	0.6098	No
Stability[d]	76.61	79.48	78.04	0.5615	No
iPro54[e]	77.76	83.15	80.45	0.6100	No
iPromoter-2L[f]	79.20	84.16	81.68	0.6343	Yes

[a]See Equation (11) for the metrics' definition.
[b]The predictor reported (Li and Lin, 2006) based on hexamers.
[c]The predictor reported (Song, 2012) with parameter $n = 4095$, $P = 600$.
[d]The predictor reported (de Avila e Silva et al., 2014) with 80, 5 and 1 neuron in input layer, hidden layer and output layer, respectively.
[e]The predictor reported in (Lin et al., 2014) with parameter $k = 7$, $\lambda = 40$, $w = 0.1$.
[f]The predictor proposed with $\xi = 8$, $\delta = 3$, K = 3, $\lambda = 2$, $w = 0.5$ and $\mathscr{F} = 1900$.
[g]Ability to identify which of the six promoter types in Equation (1).

3.3 An incisive analysis

Why can the new predictor be so powerful? To address this problem, let us consider the conservation profiles of nucleic acid residues along the DNA samples of Equation (2), as shown in Figure 2. As we can see from its panel (a), compared with the non-promoters, the profile of promoters has two distinct regions with remarkably higher conservation scores in the ranges of (−13 to −4) and (−1 to 2). Therefore, it is feasible to develop a predictor by the existing approaches to discriminate promoters from non-promoters as done by many previous investigators (see, e.g. de Avila e Silva *et al.*, 2014; Li and Lin, 2006; Lin *et al.*, 2014; Song, 2012). In the conservation profiles for the six different promoter types (Fig. 2b), however, there are many overlap regions. Therefore, it is not feasible to develop an effective predictor for identifying different types of promoters by only consider the local or short-range information with the kmer approach. In other words, some additional information is needed. Actually, in developing the iPro54-PseKNC predictor (Lin *et al.*, 2014), the authors did consider both the local and global effects with the PseKNC approach (Chen *et al.*, 2014c). That is why it can achieve better prediction quality than the other predictors as shown in Table 2. Stimulated by these authors' approach, in this study we not only considered the short- and long-range effects but also the middle-range information by introducing the multi-window-based PseKNC approach. That is the essence why the proposed predictor not only can achieve remarkably better results than the existing methods in identifying promoters but also have the ability to discriminate their types, which is beyond the reach of its cohorts.

3.4 Web server and user guide

As pointed out in (Chou and Shen, 2009), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors or any computational tools. Actually, theoretical papers with web-servers (Chen *et al.*, 2014b, 2016a, 2017; Jia *et al.*, 2016a; Liu *et al.*, 2017c; Qiu *et al.*, 2016a,c, 2017a; Xiao *et al.*, 2016; Xu *et al.*, 2014b, 2017; Zhang *et al.*, 2016) are much more attractive to broad experimental scientists (Chou, 2015). In view of this, the web-server for iPromoter-2 L has been established. Furthermore, to maximize the convenience of most experimental scientists, below is a step-by-step guide by which users can easily get their desired results without the need to go through the mathematical details.

Step 1. Click the link at http://bioinformatics.hitsz.edu.cn/iPromoter-2L/ to open the web-server iPromoter-2 L. Shown in Figure 3 is its top-page. To see a brief introduction about the predictor, click on the Read Me button.

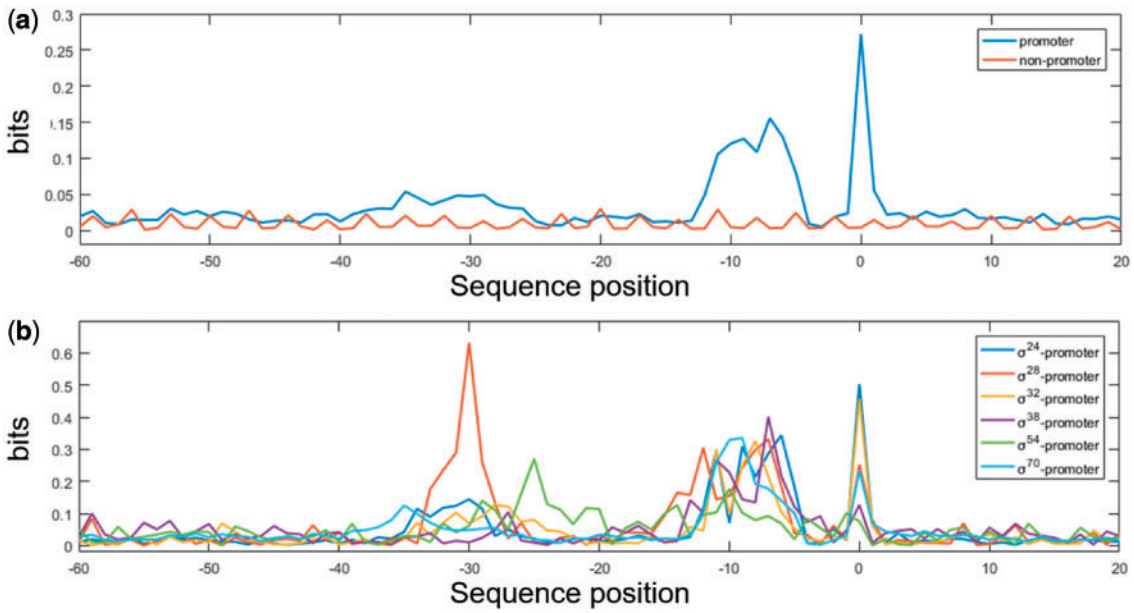

Fig. 2. The conservation profile or distribution generated by the method in (Schneider and Stephens, 1990) for **(a)** the promoter samples and non-promoter samples in $\mathbb{S}$, and **(b)** the six types of promoter samples in $\mathbb{S}^+$

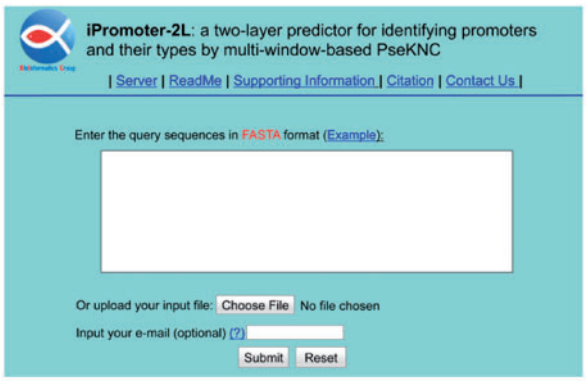

Fig. 3. A semi screenshot to show the home page of the web server iPromoter-2 L. at http://bioinformatics.hitsz.edu.cn/iPromoter-2L/

Step 2. Either type or copy/paste the query DNA sequences into the input box at the center of Figure 3. You may also upload your input via the Browse button. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the three query DNA sequences in the Example window as the input, in 30 seconds or so after submitting, you will see the following outcomes shown on the screen: (i) the first query sequence is of non-promoter; (ii) the second query sequence is of σ^{24}-promoter; (iii) the third query sequence is of σ^{70}-promoter. All these predicted results are fully consistent with experimental observations.

Step 4. Click the on Supporting Information button to download the benchmark dataset used in this study.

Step 5. Click on the Citation button to find the relevant papers that have played the key roles in developing the iPromoter-2 L predictor.

Acknowledgements

The authors wish to thank the three anonymous reviewers whose constructive comments were very helpful to strengthen the presentation of this study.

Funding

This work was supported by the National Natural Science Foundation of China [No. 61672184, 61732012, 61520106006, 31571364, U1611265], the Natural Science Foundation of Guangdong Province [2014A030313695], Guangdong Natural Science Funds for Distinguished Young Scholars [2016A030306008], Scientific Research Foundation in Shenzhen [Grant No. JCYJ20150626110425228, JCYJ20170307152201596], and Guangdong Special Support Program of Technology Young talents [2016TQ03X618].

Conflict of Interest: none declared.

References

Ahmad,K. *et al.* (2016) Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition. *J. Membr. Biol.*, **249**, 293–304.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chen,C. *et al.* (2012) Dual-layer wavelet SVM for predicting protein structural class via the general form of chou's pseudo amino acid composition. *Protein Pept. Lett.*, **19**, 422–429.

Chen,J. *et al.* (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **33**, 423–428.

Chen,W. *et al.* (2016a) iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget*, **7**, 16895–16909.

Chen,W. *et al.* (2015a) iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.

Chen,W. *et al.* (2017) iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget*, **8**, 4208–4217.

Chen,W. *et al.* (2014a) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.*, **462**, 76–83.

Chen,W. *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.

Chen,W. *et al*. (2014b) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed. Res. Int*., **2014**, 623149.

Chen,W. *et al*. (2014c) PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem*., **456**, 53–60.

Chen,W. *et al*. (2015b) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol BioSyst*., **11**, 2620–2634.

Chen,W. *et al*. (2016b) iRNA-PseU: Identifying RNA pseudouridine sites. *Molecular Therapy - Nucleic Acids*, **5**, e332.

Chen,W. *et al*. (2015c) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.

Cheng,X. *et al*. (2017a) pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. *Mol. BioSyst*., **13**, 1722–1727.

Cheng,X. *et al*. (2017b) iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, **33**, 341–346. (Corrigendum, ibid., 2017, Vol.33, 2610).

Cheng,X. *et al*. (2017c) iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget*, **8**, 58494–58503.

Cheng,X. *et al*. (2017d) pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, *Gene*. **628**, 315–321.

Cheng,X. *et al*. (2017e) pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, *Genomics*. doi:10.1016/j.ygeno.2017.08.005.

Cheng,X. *et al*. (2017f) pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, *Genomics*. doi:10.1016/j.ygeno.2017.08.005.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet*., (Erratum: ibid., 2001, Vol.44, 60). **43**, 246–255.

Chou,K.C. (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics*, **6**, 262–274.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol*., **273**, 236–247.

Chou,K.C. (2013) Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst*., **9**, 1092–1100.

Chou,K.C. (2015) Impacts of bioinformatics to medicinal chemistry. *Med. Chem*., **11**, 218–234.

Chou,K.C. (2017) An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem*., **17**, 2337–2358.

Chou,K.C. and Shen,H.B. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res*., **5**, 1888–1897.

Chou,K.C. and Shen,H.B. (2007a) MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun*., **360**, 339–345.

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

Chou,K.C. and Shen,H.B. (2007c) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun*., **357**, 633–640.

Chou,K.C. and Shen,H.B. (2009) Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci*., **01**, 63–92.

Chou,K.C. *et al*. (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One*, **6**, e18258.

Chou,K.C. *et al*. (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst*., **8**, 629–641.

Chou,K.C. and Zhang,C.T. (1995) Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*., **30**, 275–349.

de Avila e Silva,S. *et al*. (2014) DNA duplex stability as discriminative characteristic for Escherichia coli σ54- and σ28- dependent promoter sequences. *Biologicals*, **42**, 22–28.

Ding,H. *et al*. (2014) iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed. Res. Int*., **2014**, 286419.

Esmaeili,M. *et al*. (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol*., **263**, 203–209.

Feklistov,A. *et al*. (2014) Bacterial sigma factors: a historical, structural, and genomic perspective. *Annu. Revi. Microbiol*., **68**, 357–376.

Feng,P. *et al*. (2017) iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Mol. Ther. Nucleic Acids*, **7**, 155–163.

Feng,P.M. *et al*. (2013) iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem*., **442**, 118–125.

Gama-Castro,S. *et al*. (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res*., **44**, D133–D143.

Hajisharifi,Z. *et al*. (2014) Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol*., **341**, 34–40.

Han,G.S. *et al*. (2014) A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. *J. Theor. Biol*., **344**, 31–39.

Janga,S.C. and Collado-Vides,J. (2007) Structure and evolution of gene regulatory networks in microbial genomes. *Res. Microbiol*., **158**, 787–794.

Jia,J. *et al*. (2015) iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *J. Theor. Biol*., **377**, 47–56.

Jia,J. *et al*. (2016a) iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget*, **7**, 34558–34570.

Jia,J. *et al*. (2016b) Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). *J. Biomol. Struct. Dyn*., **34**, 1946–1961.

Jia,J. *et al*. (2016c) iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules*, **21**, 95.

Jia,J. *et al*. (2016d) iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem*., **497**, 48–56.

Jia,J. *et al*. (2016e) pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol*., **394**, 223–230.

Jia,J. *et al*. (2016f) pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics*, **32**, 3133–3141.

Kabir,M. and Hayat,M. (2016) iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet. Genomics*, **291**, 285–296.

Kandaswamy,K.K. *et al*. (2011) AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol*., **270**, 56–62.

Kumar,R. *et al*. (2015) Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine. *J. Theor. Biol*., **365**, 96–103.

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.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lin,H. *et al*. (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.

Lin,H. *et al*. (2017) Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, doi: 10.1109/TCBB.2017.2666141.

Lin,W.Z. *et al*. (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS One*, **6**, e24756.

Liu,B. *et al*. (2015a) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One*, **10**, e0121501.

Liu,B. *et al*. (2016a) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **32**, 362–369.

Liu,B. *et al*. (2015b) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol*., **385**, 153–159.

Liu,B. *et al*. (2015c) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.

Liu,B. *et al*. (2016b) repRNA: a web server for generating various feature vectors of RNA sequences. *Mol. Genet. Genomics*, **291**, 473–481.

Liu,B. *et al*. (2016c) iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an en-semble learning framework. *Bioinformatics*, **32**, 2411–2418.

Liu,B. *et al*. (2017a) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.

Liu,B. *et al*. (2013) Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Informatics*, **32**, 775–782.

Liu,B. *et al*. (2017b) Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences. *Nat. Sci*., **9**, 67–91.

Liu,B. *et al*. (2017c) Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. *Oncotarget*, **8**, 13338–13343.

Liu,B. *et al*. (2017d) 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Ther Nucleic Acids*, **7**, 267–277.

Liu,L.M. *et al*. (2017e) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem*., **13**, 552–559.

Liu,Z. *et al*. (2015d) iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem*., **474**, 69–77.

Liu,Z. *et al*. (2016d) pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal. Biochem*., **497**, 60–67.

Meher,P.K. *et al*. (2017) Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep*., **7**, 42362.

Mohabatkar,H. *et al*. (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine. *J. Theor. Biol*., **281**, 18–23.

Nanni,L. and Lumini,A. (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **34**, 653–660.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Potvin,E. *et al*. (2008) Sigma factors in Pseudomonas aeruginosa. *FEMS Microbiol. Rev*., **32**, 38–55.

Pugalenthi,G. *et al*. (2012) RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. *Protein Pept. Lett*., **19**, 50–56.

Qiu,W.R. *et al*. (2017a) iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. *Med. Chem*., doi: 10.2174/1573406413666170623082245.

Qiu,W.R. *et al*. (2017b) iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Informatics*, **36**, UNSP 1600010.

Qiu,W.R. *et al*. (2016a) iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget*, **7**, 44310–44321.

Qiu,W.R. *et al*. (2016b) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **32**, 3116–3123.

Qiu,W.R. *et al*. (2014) iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int J Mol Sci* (*IJMS*), **15**, 1746–1766.

Qiu,W.R. *et al*. (2016c) iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget*, **7**, 51270–51283.

Rahimi,M. *et al*. (2017) OOgenesis_Pred: a sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition. *J. Theor. Biol*., **414**, 128–136.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Shen,H.B. and Chou,K.C. (2007) Signal-3L: a 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun*., **363**, 297–303.

Song,K. (2012) Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. *Nucleic Acids Res*., **40**, 963–971.

Wang,P. *et al*. (2011) NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS One*, **6**, e23505.

Wu,Z.C. *et al*. (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst*., **7**, 3287–3297.

Xiao,X. *et al*. (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *J. Biomol. Struct. Dyn*., **33**, 2221–2233.

Xiao,X. *et al*. (2011a) GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol. Biosyst*., **7**, 911–919.

Xiao,X. *et al*. (2011b) Quat-2L: a web-server for predicting protein quaternary structural attributes. *Mol. Divers*., **15**, 149–155.

Xiao,X. *et al*. (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem*., **436**, 168–177.

Xiao,X. *et al*. (2016) iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget*, **7**, 34180–34189.

Xu,Y. *et al*. (2013a) iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **8**, e55844.

Xu,Y. *et al*. (2017) iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med. Chem*., **13**, 554–551.

Xu,Y. *et al*. (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ*., **1**, e171.

Xu,Y. *et al*. (2014a) iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci*., **15**, 7594–7610.

Xu,Y. *et al*. (2014b) iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One*, **9**, e105018.

Zhang,C.J. *et al*. (2016) iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget*, **7**, 69783–69793.

Zhang,C.T. (1997) A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol*., **187**, 297–306.

Zhang,C.T. *et al*. (2003) The Z curve database: a graphic representation of genome sequences. *Bioinformatics*, **19**, 593–599.